Proceedings

The 2009 International Symposium Computer Science and Computational Technology (ISCSCT 2009)

26 – 28, Dec. 2009

Huangshan, China

Editors

Fei Yu
Guangxue Yue
Jian Shu
Yun Liu

Co-Sponsored by

Jiaxing University, China
Peoples'Friendship University of Russia, Russia
Nanchang HangKong University, China
Feng Chia University, Taiwan
Qingdao University of Science & Technology, China
Hunan Agricultural University, China
Guangdong University of Business Studies, China
Academy Publisher, Finland

The papers in this book published by the Academy Publisher, Post: P.O. Box 40, FIN-90571, Oulu, Finland, Email: general@academypublisher.com, Internet: http://www.academypublisher.com/, Phone: +358 (0)44 525 7800, Fax: +358 (0)207 81 8199. The book is made available on-line at http://www.academypublisher.com/proc/.

AP Catalog Number AP-PROC-CS-09CN005

ISBN 978-952-5726-07-7 (Print)
978-952-5726-08-4 (CD-ROM)

Additional copies may be ordered from:

Academy Publisher
P.O. Box 40, FIN-90571
Oulu, Finland,
Phone: +358 (0)44 525 7800
Fax: +358 (0)207 81 8199
Email: order@academypublisher.com

ACADEMY PUBLISHER
http://www.academypublisher.com/

Table of Contents

Message from the Symposium Chairs

It may be said that modern information technology rests on three technology pillars: semiconductor technology, communications technology, and information processing technology. During the past two decades there have been tremendous advances in all three pillars and their unrelenting convergence that have brought us significantly closer to the realization of the long-envisioned dream of an information society in which anyone can access the cumulative knowledge of the mankind inexpensively from anywhere at any time, namely that of a ubiquitous information age. The 2009 International Symposium on Computer Science and Computational Technology will bring together researchers from academia and industry, who are interested in the emergent field of information processing. We are soliciting papers that present recent results, as well as more speculative presentations that discuss research challenges, define new applications, and propose methodologies for evaluating and the roadmap for achieving the vision of Computer Science and Computational Technology.

Welcome to ISCSCT 2009. Welcome to Huangshan, China. The 2009 International Symposium on Computer Science and Computational Technology (ISCSCT 2009) is Co-sponsored by Jiaxing University, China, Peoples'Friendship University of Russia, Russia, Nanchang HangKong University, China, Sichuan University,China, Hunan Agricultural University, China, Feng Chia University, Taiwan, Guangdong University of Business Studies, China, Academy Publisher of Finland, Finland. Much work went into preparing a program of high quality. The conference received 270 paper submissions from 17 countries and regions; every paper was reviewed by 2 program committee members; 139 papers have been selected as regular papers, representing a 51% acceptance rate for regular papers. From these 139 research papers, through two rounds of reviewing, the guest editors selected 27 papers as the Excellent papers will be published by the special issues on Journal of Computers (EI Compendex, ISSN 1796-203X), Journal of software (EI Compendex, ISSN 1796-217X), Journal of Multimedia (EI Compendex, ISSN 1796-2048), Journal of Networks (EI Compendex, ISSN 1796-2056).

The goal of ISCSCT 2009 is to bring together researchers and practitioners from academia, industry, and government to exchange their research ideas and results, and to discuss the state of the art in the many areas of security for electronic commerce. Participants of the Symposium will hear from renowned keynote speakers including: IEEE & IET Fellow Prof. Chin-Chen Chang from National Chung Hsing University, Taiwan, and Prof. Jian Shu from Nanchang HangKong University, China.

We would like to thank the program chairs, organization staff, and the members of the program committees for their hard work. We hope that ISCSCT 2009 will be successful and enjoyable to all participants.

We thank the Academy Publisher in Finland for the wonderful editorial service to this proceeding.

We wish each of you successful deliberations, stimulating discussions, new friendships and all enjoyment Huangshan, China can offer you. While this is a truly remarkable Symposium, there is more yet to come. We look forward to seeing all of you next year at the ISCSCT 2010.

Fei Yu, Guangxue Yue, Jian Shu, Yun Liu

ISCSCT 2009

Organizing Committee

Honorary Chairs

Jun Wang, *Chinese University of Hong Kong, Hong Kong (IEEE Fellow)*
Chin-Chen Chang, *National Chung Hsing University, Taiwan (IEEE & IET Fellow)*

Program Committee Chairs

Jian Shu, *Nanchang HangKong University, China*
Yongjun Chen, *Guangdong University of Business Studies, China*
Guiping Liao, *Hunan Agricultural University, China*

Organizing Chairs

Guangxue Yue, *Jiaxing University, China*
Jiexian Zeng,*Nanchang HangKong University,China*
Jun Zhang, *Guangdong University of Business Studies, China*

Finance Chairs

Fei Yu, *Peoples' Friendship University of Russia, Russia*
Yun Liu, *Qingdao University of Science & Technology, China*

Publication Chairs

Fei Yu, Peoples' Friendship University of Russia, Russia
Guangxue Yue, *Jiaxing University, China*
Jian Shu, *Nanchang HangKong University, China*
Yun Liu, *Qingdao University of Science & Technology, China*

Program Committee Members

Prof. Jianyun Nie, *University of Montreal, Canada*
Prof. Chen Xu, *Hunan University, China*
Prof. Chia-Chen Lin, *Providence University, Taiwan*
Prof. Chin-Chen Chang, *National Chung Hsing University, Taiwan*
Prof. Chu-Hsing Lin, *Tunghai University, Taiwan*
Prof. Dengyi Zhang, *Wuhan University, China*
Prof. Derong Liu, *University of Illinois at Chicago, USA*
Prof. Dongfen Yuan, *Shandong University, China*
Prof. Farong Zhong, *Zhejiang Normal University, China*
Prof. Gary G. Yen, *Oklahoma State University, USA*
Prof. Golodova Zhan Na, *Peoples' Friendship University of Russia, Russia*
Prof. Guangxue Yue, *Jiaxing University, China*
Prof. Guiping Liao, *Hunan Agricultural University, China*
Prof. Guosheng Chen, *Nanjing University of Information Science and Technology, China*
Prof. GuoZhu Liu, *Qingdao University of Science & Technology, China*
Prof. Haigang Li, *Shanghai Jiaotong University， China*
Prof. Huan Yu, *Nanchang HangKong University, China*

Prof. Hui Sun, *Nanchang Institute of Technology, China*
Prof. Jian Shu, *Nanchang HangKong University, China*
Prof. Jianghe Li, *Information Technology Department of Jiangxi Province, China*
Prof. Jie Lin, *Tongji University，China*
Prof. Jiexian Zeng, *Nanchang HangKong University, China*
Prof. Jiliu Zhou, *Sichuan University, China*
Prof. Jun Chu, *Nanchang HangKong University, China*
Prof. Jun Wang, *Chinese University of Hong Kong, Hong Kong*
Prof. Jun Zhang, *Guangdong University of Business Studies, China*
Prof. Karpus Nikolay, *Peoples' Friendship University of Russia, Russia*
Prof. Lei Shi, *Zhengzhou University, China*
Prof. Li hongquan, *Central South University, China*
Prof. Li jian, *Beijing University of Chemical Technology, China*
Prof. Limin Sun, *Institute of Software, Chinese Academy of Sciences, China*
Prof. Ming LI, *Nanchang HangKong University, China*
Prof. Mingwen Wang, *Jiangxi Normal University, China*
Prof. Mingyan Wang, *Nanchang University, China*
Prof. Naiping Hu, *Qingdao University of Science & Technology, China*
Prof. Qiang Liu, *Qingdao University of Science & Technology, China*
Prof. Qingling Li, *Qingdao University of Science & Technology, China*
Prof. Renfa Li, *Hunan University, China*
Prof. Roy Ng, *Ryerson University, Canada*
Prof. Sio-Iong Ao, *University of Oxford, UK*
Prof. Tzong-Chen Wu, *National Taiwan University of Science and Technology, Taiwan*
Prof. Weidong Zhao, *Fudan University，China*
Prof. Wen Chen, *Shanghai Jiaotong University, China*
Prof. Xiaoli Wang, *Tongji University, China*
Prof. Yongjun Chen, *Guangdong University of Business Studies, China*
Prof. Yu-Chen Hu, *Providence University, Taiwan*
Prof. Yung-Kuan Chan, *National Chung Hsing University, Taiwan*
Prof. Yuping Hu, *Guangdong University of Business Studies, China*
Prof. Youfu Du, *Yangtze University, China*

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 001-006

An Ingenious Data Hiding Scheme for Color Retinal Image

C. C. Chang[1] , Z. H. Wang[2] , and Z. X. Yin[3]
[1] Department of Information Engineering and Computer Science, Feng Chia University
Taichung 40724, Taiwan, R.O.C.
ccc@cs. ccu.edu.tw
[2] School of Software, Dalian University of Technology, Dalian, Liaoning, China
wangzhihui1017@yahoo.cn
[3] School of Computer Science and Technology, Anhui University, Hefei 230039
adyzx@qq.com

Abstract—**Steganography is the art of communicating messages in a way such that only the intended recipient knows of their existence. In this paper, we present a novel steganographic method, that can transmit the embedded secret data effectively and securely. The proposed data hiding method makes use of color retinal images as the cover media and employs the least significant bit method to embed secret bits. In essence, each color image is composed of red-green-blue planes; our scheme embeds secret bits into one of the planes of the color image to enhance the security of the transmitted message. The definition of the blood vessels plays a key role in the quality of a retinal image. More importantly, the fidelity of the blood vessel is the soul of the basis in the diagnostic procedure. How to conceal the secret data into a retinal image while introducing no or small-beer distortion to the blood vessel is one of the challenges in this research. The experimental results demonstrate that the proposed method achieves good visual quality and acceptable embedding capacity.**

Index Terms—**Data hiding; Steganography; Retinal image; LSB**

I. INTRODUCTION

Message transmission over the Internet is quite common. Some problems may arise due to security flaws in the communication channel. Steganography is the art of communicating hidden messages in a way such that no one apart from the intended recipient knows of their existence. Several steganographic schemes have been proposed recently [1-7]. The object embedded with the secret data is called a stego object. For example, a digital image with secret data is called a stego image. The goal of steganography is to communicate as many bits as possible without introducing any detectable artifacts into the stego objects. Unintended recipients of a stego object are unaware of the existence of hidden data. Because of the limitation of human senses, a small distortion in multimedia can sometimes be hard to identify. This limitation provides a space for researchers to insert the data within the multimedia. A formal definition of steganographic security can be found in [8-10].

The least significant bit (LSB) is the bit position in a binary integer. It is analogous to the least significant digit of a decimal integer, which is the digit in the ones (right-most) position.The binary representation of decimal 129(shown in Fig. 1), with the LSB in red.

1	0	0	0	0	0	0	1

Fig. 1. the least significant bit of a binary integer

The least significant bits (LSB) replacement method uses some least significant bits of cover pixels to embed secret data [11–13]. Its main advantages are simlpe implementation and low computation cost.

Here, we propose a simple information hiding scheme that can transmit the embedded secret data effectively and securely via color retinal images. The retina is important in medical treatment. Clinicians can diagnose many diseases, such as hypertension, arteriosclerosis, and blindness by observing the tortuosity changes of blood vessel in the retina. Because the background of retinal images are usually sensitive to noise and blood vessels are too important to modify, we retain those pixels without changes. Then, we embed secret messages into the color retinal image.

We generate many segmentation images of the blood vessel in different stages of the embedding and the extracting phase using the method put forth by Chang et al. [14], which is detailed in Section 2.1. From the experimental results, we conclude that the proposed method achieves good visual quality and acceptable embedding capacity.

The rest of the paper is organized as follows. To make this paper self-contained, Section 2 provides a brief introduction to related concepts and works. Section 3 presents the proposed secret embedding and extracting schemes. The experiments and results are demonstrated in Section 4. In Section 5, we make some conclusions.

II. RELATED WORKS

In this Section we review some elementary concepts and predecessor's works relevant to our study to make this paper self-contained.

A. Chang et al.'s segmentation method

Chang et al. propose an automated classification mechanism for retinal images that uses three procedures:

AP-PROC-CS-09CN005

vessel segmentation, feature reduction, and classification. In this section, we detail vessel segmentation.

Fig. 2 shows the structure of the line detector used in Chang et al.'s segmentation method. The line detector is a square window 15×15 in size. Twelve lines pass through the central pixel and the interval angle between the two lines is equipotent. It takes four steps to segment the retinal image.

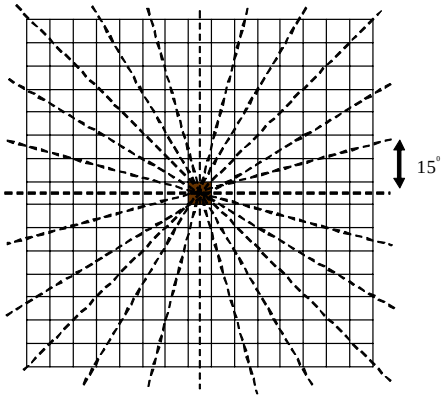


Fig. 2. The structure of the line detector

Step 1: Map the central pixel of the line detector to each pixel of the inverted green plane of the retinal image.

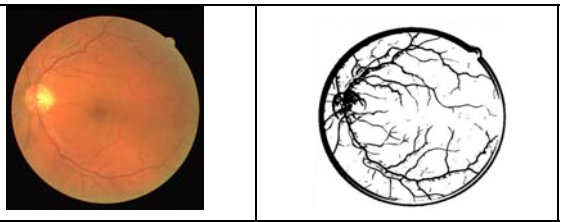
Step 2: Line detector scans the 12 lines passing through the central pixel to obtain the largest average value denoted as $L(i, j)$.

Step 3: Calculate the local average value of the 15×15 square window is calculated and denoted as $N(i, j)$.

Step 4: The difference between $L(i,j)$ and $N(i,j)$, which can be computed in Eq. (1), is output as segmentation.

$$X(i, j) = L(i, j) - N(i, j). \quad (1)$$

Complete segmentation of a retinal image is obtained after all pixels in the green plane are processed. Fig. 3 shows a retinal image and the corresponding segmentation.



(1) The retinal image	(2) The segmentation

Fig. 3. A retinal image and its segmentation

B. Least significant bits (LSB) method

Data hiding in the LSB of the pixels in a cover image is a simple algorithm with high embedding capacity. A commonly used method is LSB replacement, which is a simple way to hide information into a cover image: a great amount of bits can be embedded without causing perceptible degradation. Indeed, digital images are generally coded with eight bits using a colour channel, so insertion in the least significant bits is not visible.

As shown in Fig. 4, a cover pixel is partitioned into two bit strings: most significant and least significant. For the k-bit simple LSB replacement method, the least significant bits (i.e.,$c_k c_{k-1}...c_{1)}$ will be directly replaced by a secret string (i.e. $s_k s_{k-1}$ ...s_1) as shown:

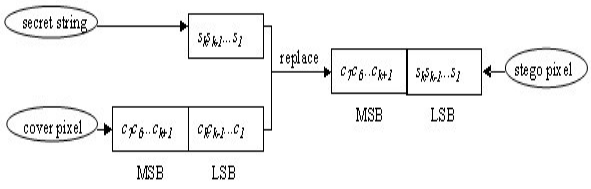


Fig. 4: k-bit simple LSBs replacement

III. THE PROPOSED SCHEME

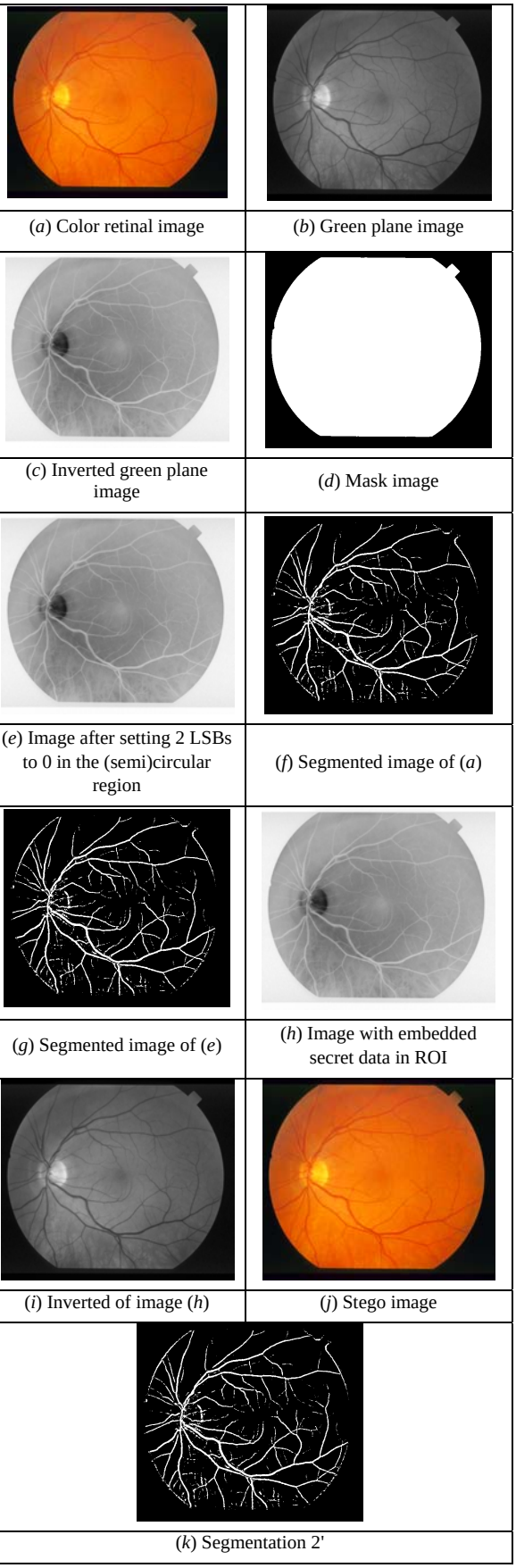
The proposed scheme is a data hiding method that makes use of color retinal images as the cover medium. Unlike the other schemes, we embed secret bits into the inverted green plane of the color image [15-17]. This makes the transmitted information more secure. The definition of the blood vessels plays a key role in the quality of a retinal image; to put it in another way, the blood vessel pixels in a color retinal image draw attention and seriously affect the doctor's diagnosis. For medical images, it is necessary to ensure that data hiding does not affect the medical value of the images. How to conceal the secret data without changing the blood vessel pixels is one of difficulties that we have to address. Moreover, how to allow the recipient to extract the hidden data (e.g. case history) quickly without too much additional information poses another challenge. To address these concerns, the proposed scheme contains two phases: embedding and extracting. We elaborate on these two phases in the following sections.

A. The embedding phase

Each color image is composed of three planes: red, green, and blue. For the human eyes, the blue plane is too dark, and the red plane is too bright. Thus, we extract the green plane and invert it as a gray-level image (shown in Figs. 5b and 4c. Since the background of the retinal images (i.e. non-color area) is usually sensitive to noise, we adopt a mask image to exclude it from further calculations. The retinal mages and the mask images (an example shown in Fig. 5d) used in this paper are from STARE database.

The blood vessel area is the key part of a retinal image. Its pixels are too sensitive and pivotal to be modified. Consequently, we embed the secret data only into the pixels belonging to the (semi)circular retinal fundus region-of-interest (ROI). This ROI excludes the background and the blood vessels. We locate the ROI as follows:

Step 1: As the frame of reference, generate segmentation 1 (Fig. 5f) and it is the original blood vessel generated from the retinal image.



(*a*) Color retinal image | (*b*) Green plane image

(*c*) Inverted green plane image | (*d*) Mask image

(*e*) Image after setting 2 LSBs to 0 in the (semi)circular region | (*f*) Segmented image of (*a*)

(*g*) Segmented image of (*e*) | (*h*) Image with embedded secret data in ROI

(*i*) Inverted of image (*h*) | (*j*) Stego image

(*k*) Segmentation 2'

Fig. 5. All the images used and generated

Step 2: Set the 2 LSBs of all pixels to 0 in the (semi)circular region of image (*c*) to obtain image (*e*).

Step 3: Generate segmentation 2 (Fig. 5g) to obtain a new blood vessel image generated from image (*e*) to use as the frame of ROI.

In this paper, ROI is defined as the region excluding

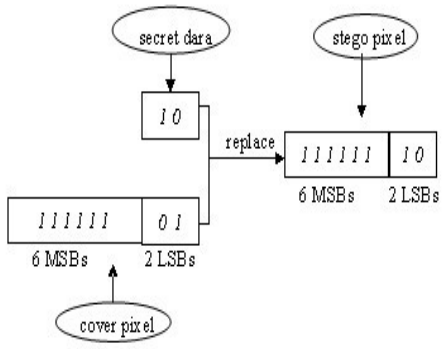


Fig. 6. Example of simple LSB replacement approach

the background, as determined by the mask image (*d*) and, the blood vessels, as determined by segmentation 2 (*g*), in the entire image. Is it doable to use segmentation 2 as the substitute of blood vessel segmentation 1 to enact ROI? We determine the feasibility using the experimental data shown in Table 1 Row 1 in Section 4. The mean value of the PSNR between segmentation 1 and segmentation 2 of the three retinal images is 22.018, which means little distortion. Thus, we embed the secret data into the ROI trustingly.

Apropos of the hiding minutia, we embed the binary secret string into the 2-rightmost least significant bits of the cover pixels. Assume that cover pixel C=253 =$(11111101)_2$, and that the secret data S=(10). The embedding operation is shown in Fig.6 and the stego pixel C'=254 =$(11111110)_2$.

When all of the secret bits are embedded, we obtain image (*h*) as shown in Fig. 5. Next, we invert image (*h*) and combine it with the blue plane and with the red plane to generate the color stego image (*j*). The entire embedding phase is described in Fig. 7.

Before transmitting the stego image, we transform the mask image into a binary string. Referring to Fig. 4d, we denote white pixels as 1 and black pixels as 0. According to the property of the mask image and the hypostasis of Run Length Encoding (RLE) [18], we compress the 0, 1 string with RLE and then send it to the recipient as a key along with the stego image.

B. The extracting phase

When the legitimate recipient receives the stego image and the information of the mask image key, he or she can extract the secret message easily in four steps:

Step 1: Generate the green plane of the received image and then invert it.

Step 2: Decompress the key to obtain the mask image.

Step 3: Using the mask image, set the two least significant bits of each pixel in the (semi)circular region of the received stego image to 0 and generate segmentation 2' (shown in Fig. 5k), which is identical with segmentation 2 to determine the ROI.

Step 4: Extract secret bits from the two least significant bits of each pixel in the ROI of the image.

Thus, the transmitted message can be extracted exactly with no additional retransmission of blood vessel segmentation. The flowchart of the extracting phase is shown in Fig. 8.

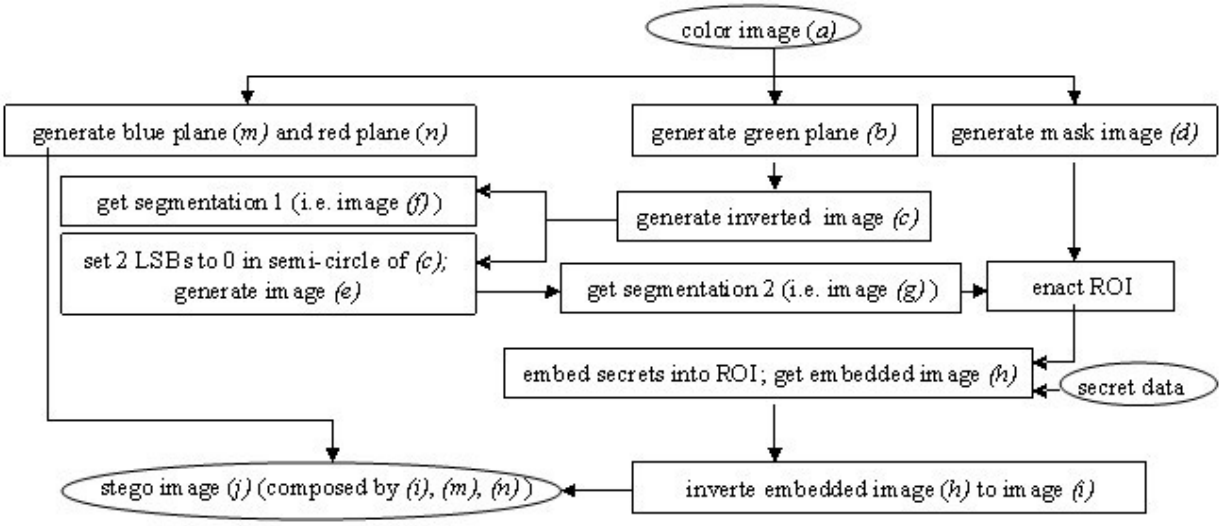


Fig. 7. Flowchart of embedding phase

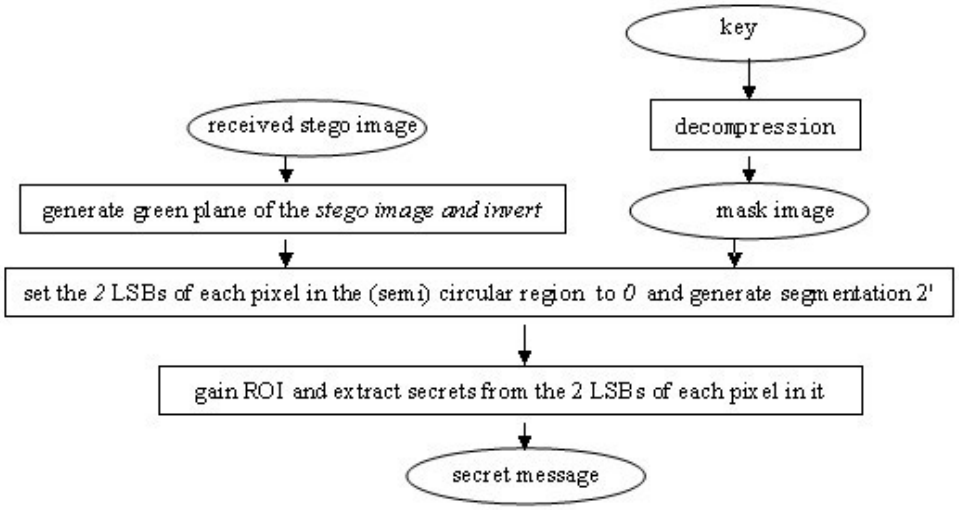


Fig. 8. Flowchart of extracting phase

IV. EXPERIMENTAL RESULTS

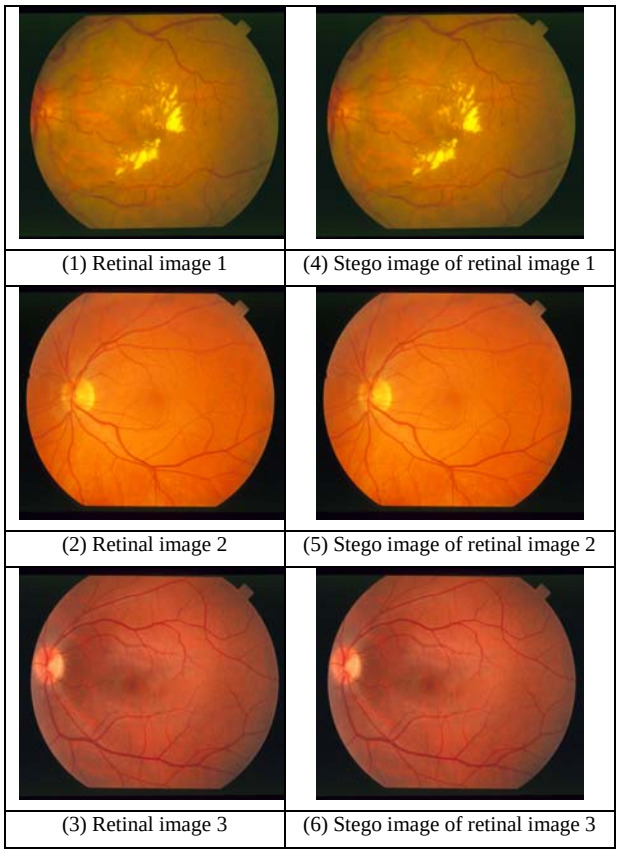
Three color retinal images of 700*605 pixels were used in the simulations of the experiments (see Fig. 9, Row 1).

The visual quality of the stego image is very important for evaluating the performance of a steganographic scheme. The human eye subjectively evaluates the visual quality of the stego image. From Fig. 8, it is difficult to distinguish the differences between the stego images and the cover images. However, judgment is influenced easily by the factors such as expertise of the viewers. To evaluate the stego image fidelity and objectively, the peak signal-to-noise ratio (PSNR), defined as Eq. (2), is adopted in this paper. PSNR is a widely used measurement for evaluating the degree of similarity between a stego image and a cover image. It can be calculated as

$$\text{PSNR}=10\log_{10}\frac{255^2}{MSE}\ (\text{dB}), \tag{2}$$

$$\text{MSE}=\frac{1}{H*W}\sum_{i=1}^{H*W}(I_i-I_i')^2, \tag{3}$$

where the mean square error (*MSE*), defined by Eq. (3) represents the difference between a stego image and its original image of $H\times W$ pixels.



(1) Retinal image 1 | (4) Stego image of retinal image 1

(2) Retinal image 2 | (5) Stego image of retinal image 2

(3) Retinal image 3 | (6) Stego image of retinal image 3

Fig. 9. Color retinal images used in the experiments and the corresponding stego images

We summarize the results of our experiments in Tables 1 and 2.

Table 1. The PSNR results of our experiments

Image from Fig. 5		Retina 1	Retina 2	Retina 3	Mean
(*f*)	(*g*)	21.465	21.834	22.755	22.018
(*a*)	(*j*)	45.532	45.324	45.565	45.474

Table2. The embedded capacity of each image

Images sized 700*605	Embedded bits (bit)	ER (bpp)
Retina 1	491174	1.160
Retina 2	494182	1.167
Retina 3	483718	1.142
Mean capacity	489691	1.156

A high *PSNR* value implies that the stego image is very similar to the original image. On the contrary, a small value of the PSNR implies that the stego image contains too many distortions and that the difference between the stego image and its original image is observable. Generally, it is difficult for the human eye to distinguish the difference between a color stego image and its original image with the *PSNR* values that are higher than 30 dB. Tables 1 and 2 show that the stego images generated by our scheme achieve very good visual quality. The PSNR mean value between the color retinal images and the corresponding stego images is 45.474 dB and the average capacity is 489691 bits (i.e. 1.156 bits per pixel).

The experimental results demonstrate that the proposed data hiding scheme achieves good visual quality and acceptable embedding capacity. Thus, we can conclude that the stego images generated by our method have a low probability of attracting attention.

V. CONCLUSIONS

In the proposed scheme, we maintain the background and blood vessel pixels of the retinal images in order to ensure the quality of the stego image. Moreover, we do not hide the secret message into the color image directly. In a creative way, we embed secret bits into the green plan of the retinal image and then reconstruct the color stego image. That is one of the originalities of our scheme distinct from the others. This enables a more secure information transfer. Additionally, we use a clever way to enact ROI (described in detail in Section 3) and then the transmitted secret data can be extracted expediently with no additional retransmission of blood vessel segmentation.

REFERRENCES

[1] Tai, W. L. and Chang, C. C., (2008):"Data Hiding Based on VQ Compressed Images Using Hamming Codes and Declustering," *International Journal of Innovative Computing, Information and Control (IJICIC)*, Vol. 5, No. 7, pp. 8-18.

[2] Chang, C. C., Lu, T. C. Chang, Y. F. and Lee, R. C. T., (2007):"Reversible Data Hiding Schemes for Deoxyribonucleic ACID (DNA) Medium," *International Journal of Innovative Computing, Information and Control (IJICIC)*, Vol. 3, No. 5, pp. 1145-1160.

[3] Chen, Y. H., Chang, C. C. and Lin, C. C., (2007): "Adaptive Data Embedding Using VQ and Clustering," *International Journal of Innovative Computing, Information and Control (IJICIC)*, Vol. 3, No. 6(A), pp. 1471-1485.

[4] Cachin, C., (1998):"An Information-Theoretic Model for Steganography," *Proc. 2nd International Workshop Information Hiding, LNCS*, Vol. 1525, pp. 306–318.

[5] Zöllner, J., Federrath, H., Klimant, H., et al., (1998):"A Novel Secret Image Sharing Scheme for True-color Images with Size Celing the Security of Steganographic Systems," *Proc. 2nd Workshop on Information Hiding, LNCS*, Vol. 1525, pp. 345-355.

[6] Katzenbeisser, S., Petitcolas F. A., (2002):"Defining Security in Steganographic Systems," *Proc. Electronic Imaging, SPIE, Security and Watermarking of Multimedia Contents IV*, Vol. 4675, pp. 50–56.

[7] Kieu, T. D., Wang, Z. H., Chang, C. C. and Li, M. C., (2009):"A Sudoku Based Wet Paper Hiding Scheme," *International Journal of Smart Home*, Vol. 3, No. 2, pp. 1-12.] Yin, Z. X., Chang, C. C., and Zhang, Y. P., (2009):"A High Embedding Efficiency Steganography Scheme for Wet Paper Codes," to appear in *Proceedings of The Fifth International Conference on Information Assurance and Security (IAS 2009)*, Xi'an, China, Aug. 18-20, 2009.

[9] Chang, C. C., Lin, C. Y. and Tseng, C. S., (2007):"Secret Image Hiding and Sharing Based on the (t, n)-Threshold," *Fundamenta Informaticae*, Vol. 76, No. 4, pp. 399-411.

[10] Luo, H., Yu, F. X., Chu, S. C., Lu, Z. M., (2007):"Hiding Multiple Watermarks in Transparencies of Visual Criptography," *International Journal of Innovative Computing, Information and Control (IJICIC)*, Vol. 5, No. 7, pp. 1875-1881.

[11] Dumitrescu, S., Wu, X., Wang, Z., (2003): " Detection of LSB Steganography via Sample Pair Analysis," *IEEE Trans. Signal Process*, Vol. 51, pp. 355-372.

[12] Ker, A., (2004):" Improved Detection of LSB Steganography in Grayscale Images," *Proc. The 6th Information Hiding Workshop*, Vol. 3200, pp. 97-115.

[13] Luo, X. Y., Liu, F. L., (2007): "A LSB Steganography Approach Against Pixels Sample Pairs Steganalysis," *International Journal of Innovative Computing, Information and Control (IJICIC)*, Vol. 3, No. 3, pp. 575-588.

[14] Chang, C. C., Chen, Y. C. and Lin, C. C., (2009): "A New Classification Mechanism for Retinal Images," *to apper in Proceedings of the International Conference on Information Technology and Computer Science*, Kiev, Ukraine, Jul. 2009.

[15] Chang, C. C., Lin C. Y., Fan Y. H., (2008): "Lossless Data Hiding for Color Images Based on Block Truncation Coding". *Pattern Recognition*, Vol. 41, pp. 2347-2357.

[16] Tsai, Y. Y., Wang C. M., (2007): "A Novel Data Hiding Scheme for Color Images Using a BSP Tree

". *The Journal of Systems and Software*, Vol. 80, pp. 429–437.

[17] Tsai, D. S., Horng G. B., Chen T. H., Huang Y. T., (2009): " A Novel Secret Image Sharing Scheme for True-color Images with Size Constraint". *Information Sciences*, Vol. 179, pp. 3247–3254.

[18] Pu, I. M., (2005): " Run-length Algorithms". *Fundamental Data Compression*, pp. 49–65.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 007-011

Key Frame Extraction from MPEG Video Stream

Guozhu Liu, and Junming Zhao
College of Information Science & Technology, Qingdao Univ. of Science & Technology,266061,P. R. China
Email: lgz_0228@163.com

***Abstract*—In order to extract valid information from video, process video data efficiently, and reduce the transfer stress of network, more and more attention is being paid to the video processing technology. The amount of data in video processing is significantly reduced by using video segmentation and key-frame extraction. So, these two technologies have gradually become the focus of research. With the features of MPEG compressed video stream, a new method is presented for extracting key frames. Firstly, an improved histogram matching method is used for video segmentation. Secondly, the key frames are extracted utilizing the features of I-frame, P-frame and B-frame for each sub-lens. Fidelity and compression ratio are used to measure the validity of the method. Experimental results show that the extracted key frames can summarize the salient content of the video and the method is of good feasibility, high efficiency, and high robustness.**

***Index Item*—MPEG coding; key frame; shot segmentation; histogram; fidelity**

I. INTRODUCTION

In order to reduce the transfer stress in network and invalid information transmission, the transmission, storage and management techniques of video information become more and more important.

Video segmentation and key frame extraction are the bases of video analysis and content-based video retrieval. Key frame extraction [1-2], is an essential part in video analysis and management, providing a suitable video summarization for video indexing, browsing and retrieval. The use of key frames reduces the amount of data required in video indexing and provides the framework for dealing with the video content [3].

In recent years, many algorithms of key frame extraction focused on original video stream. It can introduce processing inefficiency and computational complexity when decompression is required before video processing.

Key frame is the frame which can represent the salient content and information of the shot. The key frames extracted must summarize the characteristics of the video, and the image characteristics of a video can be tracked by all the key frames in time sequence. Furthermore, the content of the video can be recognized. A basic rule of key frame extraction is that key frame extraction would rather be wrong than not enough. So it is necessary to discard the frames with repetitive or redundant information during the extraction.

A new algorithm of key frame extraction from compressed video data is presented in this paper. We analyze the features of compressed data and finally obtain the key frames.

For video, a common first step is to segment the videos into temporal "shots," each representing an event or continuous sequence of actions. A shot represents a sequence of frames captured from a unique and continuous record from a camera. Then key frames are to be extracted. Video segmentation is the premise of key frame extraction, and key frames are the salient content of the video (key factors to describe the video contents). Figure 1 illustrates the basic framework of our algorithm.

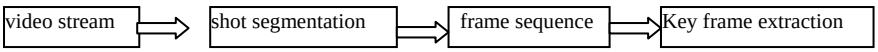


Fig. 1 The basic framework of the key frame extraction algorithm from MPEG video stream

II. ALGORITHM REALIZATIONS

A. SHOT SEGMENTATION

Shot segmentation is the first step of the key frame extraction, which mainly refers to detecting the transition between successive shots. The domain of video shot segmentation falls into two categories: uncompressed and compressed. The detection methods can be broadly classified into abrupt transition detection and gradual transition detection [4].

Nowadays, the common used methods of shot transition detection are as follows: pixel-based comparison, template matching and histogram-based method [5-6]. The pixel-based methods are highly sensitive to motion of objects. So it is suitable to detect obvious segmentation transition of the camera and object movement. Template matching is apt to result in error detection if you only simply use this method. In contrast to pixel-based methods, the Histogram-based methods completely lose the location information of pixels. Consequently, two images with similar histograms may have completely different content. In addition, there exist other methods of shot detection, such as boundary detection and dual threshold method.

When camera switches, the video image data will undergo a series of significant changes, such as content change, color difference increase and trajectory

discontinuities. Thus there is a peak between consecutive frames. A color histogram method is adopted to segment the shots according to the frame difference.

The Histogram-based method is the most common used method to calculate frame difference. Since color histograms do not relate spatial information with the pixels of a given color, and only records the amount of color information, images with similar color histograms can have dramatically different appearances. To solve the problem, an improved histogram algorithm, X^2 histogram matching method is adopted. The color histogram difference $d(I_i, I_j)$ between two consecutive frames I_i and I_j can be calculated as follows:

$$d\left(I_i, I_j\right) = \sum_{k=1}^{n} \frac{(H_{ik} - H_{jk})^2}{H_{jk} + H_{ik}}, (H_{jk} \neq 0) \quad (1)$$

Where H_i and H_j stand for the histogram of I_i and I_j, respectively.

A shot transition occurs when $d(I_i, I_j)$ is bigger than a given threshold. The experiment result illustrates that good effect can be achieved. Selecting an appropriate threshold is the key to the method.

B. Key frame extraction

Since key frame extraction plays an important role in video retrieval and video indexing, a lot of research has been done on the techniques. The widely used key frame extraction techniques are as follows:

(1) Key frame extraction based on shot activity. Gresle and Huang [7] computed the intra and reference histograms and then compute an activity indicator. Based on the activity curve, the local minima are selected as the key frames.

(2) Key frame extraction based on macro-block statistical characteristics of MPEG video stream. Janko and Ebroul [8] generate the frame difference metrics by analyzing statistics of the macro-block features extracted from the MPEG compressed stream. The key-frame extraction method is implemented using difference metrics curve simplification by discrete contour evolution algorithm.

(3) Key frame extraction based on motion analysis. Wolf [9] computed the optical flow for each frame and then used a simple motion metric to evaluate the changes in the optical flow along the sequence. Key frames are then found at places where the metric as a function of time has its local minima.

Nowadays, most of the video are stored in the compressed form of MPEG. The MPEG video compression algorithm has two main advantages: macro block-based motion compensation for the reduction of the temporal redundancy and transform domain based compression for the reduction of spatial redundancy [10]. In the compression of the video stream, frames can be grouped into sequences called a group of pictures (GOP). The types of frames can be classified into I frames, P frames and B frame. They are regularly arranged in the video stream and compose the GOPs. Figure 2 shows how different types of frames can compose a GOP.

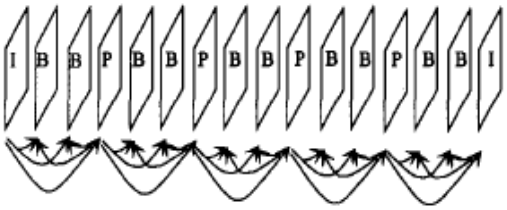


Fig. 2 The frame sequence of a MPEG GOP

Within a GOP, an I frame is the first frame and I frames and P frames are reference frames. I frames are intra-coded. The frames are processed with discrete cosine transform (DCT) using 8*8 blocks, and DC coefficients contain the main information. P frames and B frames are inter-frame coded. P frames refer to the preceding I frame or P frame, and are predicatively coded with only forward motion compensation based on macro blocks. The forward motion vectors for forward motion predication and DCT coefficients of residual error after motion compensation are obtained. B frames are inter-frame coded for forward motion prediction, backward motion prediction and bi-directional motion prediction. Each Macro-block of 16*16 pixels in P frames and B frames search for the optimal matching macro block in corresponding reference frames, then reduce predictive error of motion compensation with DCT coding. At the same time, one or two motion vectors are transferred.

Key frames are extracted using the characteristics of I frames, P frames and B frames in the MPEG video stream after shot segmentation.

If a scene cut occurs, the first I frame is chosen as a key frame.

In video stream, P frames are coded with forward motion compensation. When a shot transition occurs at a P frame, great change can take place in the P frame corresponding to the previous reference frames. So encoder can not utilize the macro blocks in previous reference frames to compensate the effect. Therefore, many of the macro blocks should have been coded as "intra" without motion compensation.

An equation is designed to calculate the ratio of macro blocks without motion compensation, which is used to detect whether the P frame is selected as a key frame. The equation is given below:

$$R_p = \frac{no\_com}{com} \quad (2)$$

Where no_com denotes the number of macro blocks without motion compensation, and com stands for the number of macro blocks after motion compensation.

When R_p peak appears, the P frame can be selected as a key frame.

Shot transition in video stream can also occur at B frame. In the circumstance, great change may take place on the content of B frame compared with the preceding frame. Therefore, the motion vectors come from the

reference frames after the B frame instead of the former ones when B frames are coded by the encoder.

A ratio of backward motion vectors and forward motion vectors is calculated to detect whether the B frame is a key frame. The equation is given as follows:

$$R_B = \frac{back}{forw} \qquad (3)$$

where $back$ is the number of the backward motion vectors and $forw$ denotes the number of the forward motion vectors.

III. VALIDITY MEASURES

Key frame extraction aims to reduce the amount of video data, and the frame sequence must preserve the overall contents of the original video. Whether the key frames can be accurately detected and extraction is the fundamental rule to measure the validity of the algorithm. Current measurement mainly relies on eye observation. If we simply use this method, it is time-consuming in the case of large-scale video data.

Compression ratio and fidelity[11-12] are chosen to measure the validity of the algorithm in our work. Compression ratio measure is used to evaluate the compactness of the key frame sequence, while fidelity is to measure the correlation degrees of the sets in the image classifications. Fidelity is defined as a Semi-Hausdorff distance between the key frame set and the shot frame set.

Suppose that the key frame set R consists of K frames, $R = \{KF_j \mid j = 1,2,\cdots,k\}$, while the shot frame set S consists of N frames, $S = \{F_i \mid i = 1,2,\cdots,k\}$. Let the distance between any two frames KF_j and F_i be $d(KF_j, F_i)$. Define d_i for each frame F_i as:

$$d_i = min(d(KF_j, F_i)),\ j = 1,2,\cdots,k \qquad (4)$$

Then the Semi-Hausdorff distance between S and R is given as:

$$d_{sh} = max(d_i),\ i = 1,2,\cdots,N. \qquad (5)$$

The fidelity measure is defined as:

$$fidelity = 1 - \frac{d_{sh}}{\max_i(\max_j(d_{ij}))} \qquad (6)$$

where d_{ij} denotes the dissimilarity matrix of the shot set S.

The bigger the fidelity is, the more accurate the global scan of key frames over the original video is.

IV. EXPERIMENT TEST

A segment of video is selected to do the experiment. Firstly, let the first frame as a key frame, and the ratios are calculated according to equation (2),(3). The frame where a ratio peak occurs is extracted as a key frame. If there are no transitions in a shot, the frames in the shot have high similarity, and there is no significant change among the characteristic curves. Then the first frame can be extracted as a key frame, and finally the key frames of the video can be obtained. The experimental result is as follows:

Fig. 3 The segment of the video

Fig. 4 Key frames extracted (the first frame, the 10th frame and the 31st frame

The algorithm is based on shot. There is only one shot in the tested video sequence, and there are no gradual transitions and abrupt transitions. So only a shot is obtained after the shot segmentation. The first frame is extracted as a key frame, and other key frames are calculated according to macro block motion when abrupt transitions occur.

Ⅴ. RESULTS AND COMPARATIVE ANALYSIS

To verify the validity of the algorithm, it is realized by VC++ in Window XP environment on Intel PD (2.80GHz) with 1G storage. Four segments of video with different characteristics are selected from The Open Video Project as training samples. Experimental test is processed using the method. The results are shown in Table 1.

TABLE 1. THE RESULTS OF THE EXPERIMENT

video sequences	number of the shot	total frames	key frames	compression ratio（%）	fidelity
A:BOR10_013	8	1026	12	99.3	0.7424
B:HURR003	12	1653	22	99.5	0.7823
C:Indi009	17	2896	30	99.1	0.7732
D:aircrash	10	1952	15	99.6	0.7534

From Table 1, the average compression ratio of the new algorithm is 99.48%, and the average fidelity is 0.76. From the experiment, we can see that the representative key frames can be extracted accurately and semantically from long video sequences or videos with more transitions, reflecting the video content objectively.

Figure 5 shows the results of key frame extraction with the proposed algorithm. The video depicts the process of plane take-off, failure, pilot parachuting. Ten shots are segmented and 15 key frames are extracted from the video with the algorithm. From the key frames, the video content can be clearly acknowledged, including the people, the process of the accident, etc. The result illustrates that the algorithm is valid to segment the shot and extract the key frames and it is of good feasibility and strong robustness.

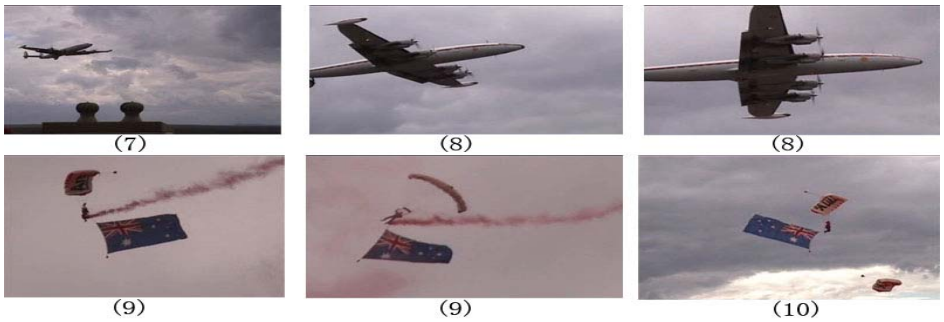


Fig.5 The results of the key frame extraction from the video sequence of air crash

In a second test, we examine the validity of our algorithm comparing the fidelity with two newer algorithms. One algorithm is the key frame extraction combining global and local information[13], and the other is information theory based shot cut/ fade detection and video Summarization[14]. The fidelity values are calculated with the 4 videos above respectively, as illustrated in Fig. 6.

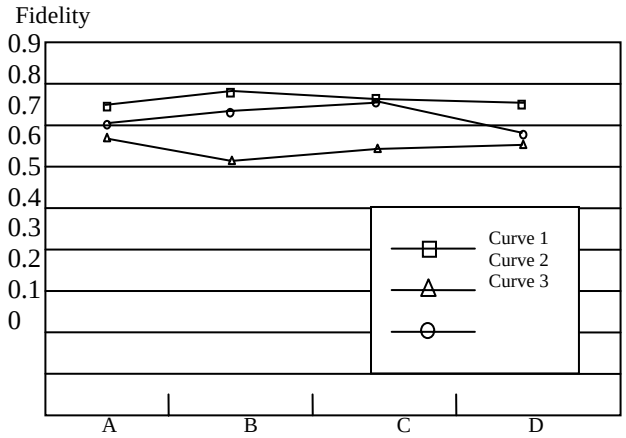


Fig. 6 Fidelity values of the three algorithms

Curve 1 indicates the results of our method.

Curve 2 shows the results of the algorithm combining the global and local information, where the parameter values are as follows: $\delta_1 = 0.4$, $\delta_2 = 0.04$, $m = 3$, $\delta = 0.8$, $dis = 13$ and $times = 5$. The algorithm adopts similarity model base on color characteristics. The fidelity value is low for the video B with simple color. At the same time, the algorithm can not achieve good effect for the longer video C. Therefore, it depends on the accuracy of the shot segmentation. Moreover, it also relies on the threshold. When the parameters change, the effects also change accordingly.

Curve 3 is the result of the last algorithm. Good results can be achieved by selecting different thresholds for different video. However, it also depends on the thresholds. Satisfactory results only can be achieved through many experiments.

From the above analysis, the algorithm we proposed is of high accuracy, good feasibility and generality.

Ⅵ. CONCLUSIONS

We have proposed a new algorithm for key frame extraction. It compensates for the shortcomings of other algorithm and improves the techniques of key frame extraction based on MPEG video stream. The experimental results show that good fidelity and compression ratio can be achieved. It is not only of good feasibility, high efficiency, but also with low error and high robustness.

REFERENCE

[1] D.Feng, W.Siu and H. Zhang, “Multimedia information retrieval and management: Technological Fundamentals and Applications, ” *Springer*, pp.44, 2003.

[2] Costas Cotsaces, Nikos Nikolaidis, and Ioannis Pitas, "Video shot detection and condensed representation: a review," I*EEE Signal Processing*, vol. 23, no. 2, pp. 28-37, 2006.
[3] T. Liu, H. Zhang, and F. Qi, "A novel video key-frame-extraction algorithm based on perceived motion energy model," *IEEE Transactions On Circuits And Systems For Video Technology*, vol. 13, no. 10, pp. 1006-1013, 2003.
[4] Irena Koprinska, Sergio Carrato, "Temporal video segmentation: A survey," *Signal Processing: Image Communication*, vol. 16, no. 5, pp. 477-500,2001.
[5] C. F. Lam, M. C. Lee, "Video segmentation using color difference histogram," *Lecture Notes in Computer Science*, New York: Springer Press, pp. 159–174., 1998.
[6] A. Hampapur, R. Jain, and T. Weymouth, "Production model based digital video segmentation," *Multimedia Tools Application*, vol. 1, no. 1, pp.9–46, 1995.
[7]P. Gresle, T. S. Huang, "Gisting of video documents: a key frames selection algorithm using relative activity measure," *The 2nd International Conference On Visual Information System*, 1997.
[8]J. Calic, E. Izquierdo, "Efficient key-frame extraction and video analysis information technology: coding and computing," *international symposium on information technology*, pp. 28-33, 2002.
[9]W. Wolf, "Key frame selection by motion analysis," *Proc. IEEE Int. Conf. Acoust., Speech Signal Proc*., vol. 2, pp. 1228–1231, 1996.
[10] D. Le Gall, "MPEG: a video compression standard for multimedia applications," *Communications of the ACM*, vol. 34, pp. 46-58, 1991.
[11] D. Besiris, N. Laskaris, F. Fotopoulou, et al., "Key frame extraction in video sequences: a vantage points approach," *2007 International Workshop on Multime*dia *Signal*, pp. 434-437, 2007.
[12] G. Ciocca and R. Schettini, "An innovative algorithm for key frame extraction in video summarization," *J. Real-Time Image Process*, vol.1, no. 1, pp. 69–88, 2006.
[13]Z. Zhan, W. Yu, "A method of key frame extraction combing global and local information," *Application Research of Computers*, vol.24, no. 11, pp. 1-4, 2007. (In Chinese)
[14] Z. Cernekova, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 82–91, Jan. 2006

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 012-015

An Algorithm for Mining Maximum Frequent Itemsets Using Data-sets Condensing and Intersection Pruning

Shui Wang[1], Ying Zhan[1], and Le Wang[2]
[1]Software School, Nanyang Institute of Technology, Nanyang, China
E-mail: seawan@163.com, abulaabula@163.com
[2]Institute of Information Engineering, Dalian University, Dalian, China
E-mail: wangleboro@163.com

***Abstract*—Discovering maximal frequent itemset is a key issue in data mining; the Apriori-like algorithms use candidate itemsets generating/testing method, but this approach is highly time-consuming. To look for an algorithm that can avoid the generating of vast volume of candidate itemsets, nor the generating of frequent pattern tree, DCIP algorithm uses data-set condensing and intersection pruning to find the maximal frequent itemset. The condensing process is performed by deleting items in infrequent 1-itemset and merging duplicate transactions repeatedly; the pruning process is performed by generating intersections of transactions and deleting unneeded subsets recursively. This algorithm differs from all classical maximal frequent itemset discovering algorithms; experiments show that this algorithm is valid with moderate efficiency; it is also easy to code for use in KDD applications.**

***Index Terms*—data mining, maximum frequent itemsets, candidate itemsets, intersection pruning, data-set condensing**

I. INTRODUCTION

Mining frequent itemsets (or patterns) is a key problem in data mining, and it's widely used in applications concerning association rules and sequential sampling models [1]. Because all frequent itemsets are considered implicitly in the maximum frequent itemset (MFI), the issue of discovering frequent itemset can be converted to the issue of discovering maximal frequent itemset. Besides, only maximal frequent itemset is needed in some of the data mining applications instead of the frequent itemset [2].

In 1998, Bayardo presented an algorithm of mining maximal frequent itemset denoted as Max-Miner[3], which used set-enumeration tree as the concept framework and adopted breadth-first searching method, as well as superset pruning strategy and dynamic recording technology (ascending sort according to itemset's support). There is another MFI algorithm, called Pincer-Search[4],which used bidirectional searching method of both top-down and bottom-top search. The Depth-Project algorithm uses sequential dictionary tree of itemsets with a depth-first approach, together with superset pruning strategy & dynamic recording technology, to search MFI [5]. In 2001, Burdick proposed an algorithm called MAFIA, which used itemset grid (similar to enumeration tree) and subset tree as concept framework [6], and stored data-set in longitudinal bitmaps. MAFIA used multiply pruning strategy including PEP (parent equivalence pruning), FHUT (frequent head union tail) and HUTMFI (using known MFI for pruning). GenMax combined the pruning and mining process [7], utilizing two strategies for MFI: one projected the found MFI to current nodes to provide fast superset checking; another used Diffset spreading for fast support calculation. In 2003, SONG Yu-qing etc. proposed an algorithm for mining & updating MFI based on FP-tree [8]; according to the mechanism of FP-tree, MFI is mapped as a unique path of the FP-tree, so the calculation of MFI support is converted to a process of calculating the count of MFI paths. In 2004, a parallel MFI algorithm was also proposed by LI Qing-hua, etc [9].

Although new algorithms are proposed constantly, but up to now, improving the time-efficiency of mining is still a key research field; and because most algorithms are based on FP-tree, they are usually hard to implement. This paper proposes an easy coding algorithm called DCIP, which uses data-set condensing and intersection pruning for mining MFI.

II. BASIC CONCEPTS

Let $I=\{i_1,i_2,...,i_m\}$ be a set of m distinct items. A transaction T is defined as any subset of items in I. A transaction database D is a set of transactions like T. A transaction T is said to support an itemset $X \subseteq I$ if it contains all items of X. The fraction of the transaction in D that support X is called the support of X, denoted as support(X). An itemset is frequent if its support is above some users defined minimum support threshold. Otherwise, it is infrequent.

Definition 1: The number of items in an itemset is called the *length* of an itemset. Itemsets of some length *k* are referred to as *k*-itemsets.

Definition 2: A frequent itemset is called maximal if it is not a subset of any other frequent itemset. The set of all maximal frequent itemsets is denoted as MFI.

So obviously, all frequent itemsets are subsets of MFI, and the discovering of frequent itemsets is converted to

the discovering of MFI

Property 1: If an itemset is infrequent, all its supersets must be infrequent.

Property 2: If an itemset is frequent, all its subsets must be frequent.

Property 3: Maximal frequent itemset is a subset of frequent itemsets.

Property 4: if Y is a subset of X (i.e. $Y \subseteq X$), then support(Y)$\geqslant$support(X).

According to the properties, we can get the following inference:

Inference 1: Assume $T=\{x_1, x_2, ..., x_{m-1}, x_m\}$, $TJ=\{x_1, x_2, ..., x_{j-1}, x_{j+1}, ..., x_{m-1}, x_m\}$, if $\{x_j\}(1 \leqslant j \leqslant m)$ is an infrequent 1-itemset, and $\{x_i\}(i=1, 2, ..., m, i \neq j)$ is a frequent 1-itemset, itemset T must be infrequent, but itemset TJ may be a MFI.

Inference 1 describes the fact that deleting items in infrequent 1-itemsets from the transaction, the set of the remaining items may be a MFI.

Lemma 1: Assume data-set $D= \{T_1, T_2, ..., T_n\}$, and DL is its maximal frequent itemsets, if minimum support threshold is s, L_i (i=1, 2, ..., n) is the maximal frequent itemset corresponding to items in T_i respectively, then $DL \subseteq L_1 \cup L_2 \cup ... \cup L_{n-s+1.}$

Proof: Any itemset of DL must be maximal frequent itemset corresponding to one transaction in D, so $DL \subseteq L_1 \cup L_2 \cup ... \cup L_n$. For any itemset $I \subseteq L_{n-s+2}$, there must be at least s transactions that are its supersets, and at least one transaction T_i $(1 \leqslant i \leqslant n-s+1)$, so $I \in L_i$ $(1 \leqslant i \leqslant n-s+1)$, and $L_{n-s+2} \subseteq L_1 \cup L_2 \cup ... \cup L_{n-s+1}$. The same reasoning process can be applied to deduce $L_{n-s+i} \subseteq L_1 \cup L_2 \cup ... \cup L_{n-s+1}$ $(2 \leqslant i \leqslant s)$, so $DL \subseteq L_1 \cup L_2 \cup ... \cup L_{n-s+1}$.

III. DCIP ALGORITHM

The first step of DCIP algorithm is to reduce the length of itemsets and the volume of data-set. According to Lemma 1, any maximal frequent itemset is also a maximal frequent itemset corresponding to one transaction in D, so find all maximal frequent itemsets correspond to every transaction through intersection pruning, merge them into one set (denoted as FS hereinafter), then delete all infrequent maximal itemset in FS, and the remaining set is maximal frequent itemset. The two main processes are described as follows.

A. Condensing the Data-set

This process first sorts the data-set with descending order according to the length of its itemsets, then moves those high-dimensional transactions whose support are bigger than minimal support threshold to a frequent itemset, and deletes all subsets of those transactions to condense the data-set. These steps are as follows:

Step 1: Scan the data-set, finding all frequent 1-itemset;

Step 2: Scan the data-set, deleting all items infrequent 1-itemset from all transactions; then add up identical transactions (i.e., if transaction T1=T2, let support(T1) = support(T1) + support(T2), and delete T2 from data-sets). Sorting the data-set descendingly according to the length of itemsets to form a new data-set which we denote as C;

Step 3: Process every transaction Ti in C whose support are bigger than minimal support threshold: move T_i to FS and delete all T_j ($T_j \subset T_i$, j>i);

Step 4: Delete non-MFI from FS;

Step 5: End.

B. Intersection Pruning

Any maximal frequent itemset is also the maximal frequent itemset corresponding to a certain transaction in D; merge all maximal frequent itemset corresponding to every transaction into one set (which we denote as FS), then delete all non-frequent maximal itemsets in FS, and the remaining set is the maximal frequent itemset. These steps are as follows:

Assume we have a data-set denoted as D, and the minimal support threshold is S.

Step 1: Condense data-set D using the method described in 3.1; if |D|<S, terminate the processing for the current data-set;

Step 2: Find intersection of T_1 and $T_i(1<i \leqslant n)$; merge all intersections into a new data-set D1; establish the vertical data format of D; delete transaction T_i ($T_i \subset T_1$); if $|D1| \geqslant S$, then go to step 1 to perform another intersection pruning circle for D1;

Step 3: Use the vertical data format of D to find the intersection of T_j and T_i (j=2, 3, 4, ..., m<n; $j<i \leqslant n$), merge all intersections into a new data-set D1, go to step 1 to perform another intersection pruning circle for D1; when the volume of the remaining data-set is less than S, stop finding intersections of T_j and T_i, terminate the process for current data-set.

Step 4: End;

Note: Data-set condensing can be performed at the beginning of the intersection pruning process, as well as in the process of step 3.

C. Instance Analysis

The following example shows how to discover MFI using DCIP for transaction database D (Table I) with minimum support threshold as 4 (i.e., minsup=4).

TABLE I.
TRANSACTION DATA-SET D

TID	Items
001	I1, I2, I4, I5, I7
002	I1, I2, I5, I6, I7
003	I0, I3, I5, I7
004	I0, I3, I8
005	I1, I2, I3, I4, I7
006	I2, I3, I7, I8
007	I3, I6, I9
008	I1, I3, I5, I9
009	I1, I2, I6
010	I3, I4, I8, I9

Step 1: Condense transaction data-set D using the method in 3.1, the result is shown in Table II;

TABLE II.
RESULT OF CONDENSED D

TID	items	Count	del
1	I1, I2, I5, I7	2	
2	I1, I2, I3, I7	1	
3	I3, I5, I7	1	
4	I2, I3, I7	1	1
5	I1, I3, I5	1	
6	I1, I2	1	1

Attribute Count is the count of corresponding transactions; attribute del indicate whether the corresponding transaction can be ignored in later processing, for example, after step 2, T6 can be ignored.

Step 2: Find intersections of T_1 and T_i (i=2, 3, ..., 7), merge all intersections into data-set D1, as shown in Table III:

TABLE III.
INTERSECTION DATA-SET FOR T1 IN TABLE II

TID	items	Count	del
1	I1, I2, I7	1(+2)	
2	I5, I7	1(+2)	
3	I2, I7	1(+2)	1
4	I1, I5	1(+2)	
5	I1, I2	1(+2)	1

Establish vertical data format for D; because in Table II, $T_6 \subset T_1$, T_6.del=1 (see Table II). The (+2) for attribute Count in table III is the count of T_1 in Table II.

Step 3: Condense the data-sets in Table III; as this example, the result remains no change.

Step 4: Find intersections of T_1 and T_i (i=2, 3, 4, 5) in Table III respectively, merge them into a new data-set D1, as shown in Table IV.

TABLE IV.
INTERSECTION DATA-SET FOR T1 IN TABLE III

TID	items	Count	del
1	I2, I7	1(+2+1)	
2	I1, I2	1(+2+1)	

Because T_3 and T_5 are subset of T_1 in Table III, delete T_3 and T_5;

Step 5: Condense the data-set in Table IV, produce frequent itemset {{I2, I7}:4, {I1, I2}:4}; Table IV is now empty after condensing;

Step 6: Back to Table III, T_3 and T_5 has been deleted, we only need to find the intersection of T_2 and T_4; but the length of T_2 and T_4 are both 2, no need to find intersection of them.

Step 7: Back to Table II, because T_6 has been deleted, we only need to find the intersections of T_2 and T_i (i=3,4,5,7); merge all intersections into a new data-set D1, as shown in Table V.

TABLE V.
INTERSECTION DATA-SET FOR T2 IN TABLE 2

TID	items	Count	del
1	I3, I7	1(+1)	
2	I2, I3, I7	1(+1)	
3	I1, I3	1(+1)	

Because $T_4 \subset T_2$ in Table II, it should be deleted.

Step 8: Condense the data-set in Table V; after condensing the result is empty;

Step 9: The original data-set D has 10 transactions; Table II shows that 30% (3 transactions) of them has been processed; condense again the remaining data-set in Table II, and the result is empty. The process ends.

Step 10: Merge all resulting frequent itemsets, and delete all non-frequent maximal itemsets, the final result of MFI is {{I2, I7}: 4, {I1, I2}: 4}.

The steps above use 14 times of intersection calculations for MFI; compared with other Apriori-like algorithms, its simplicity and efficiency is explicit.

Note: Because the volume of the example data-set D is small (only 10), the above process does not include the utilizing of vertical data format; the reason of introducing vertical data format is to reduce the number of times of finding the intersections.

IV. PERFORMANCE STUDY

If the length of the longest transaction itemset is L, the depth of recursive calling of the algorithm itself is less than L-2. The number of calculations for intersections is negatively correlated with support, number of deleted transactions, and number of duplicated transactions. This algorithm can also be implemented parallelly for each transaction's maximal frequent itemset to get more efficiency. It is valid for both long and short frequent pattern mining applications; for vast volume of data-set, its usability retains because of the time & space cost increases not very drastically. The validity of this algorithm can be assured by Theorem 1 and Lemma 1.

Another advantage of DCIP algorithm is its easy implementation. It is coded and tested using PowerBuilder script language on a microcomputer with Pentium IV/1.80GHz CPU, 512M memory running Windows XP operation system. Testing dataset is extracted from a supermarket's sales record.3000, 5000, 10000 and 20000 transactions are tested respectively with each record having 2-10 categories of commodity (the average number of categories is 6). Figure 1 shows the running time for different volume of datasets with minimum support threshold of 5%, 20% and 50%respectively. The bigger minimum support threshold, the lesser time needed for MFI.

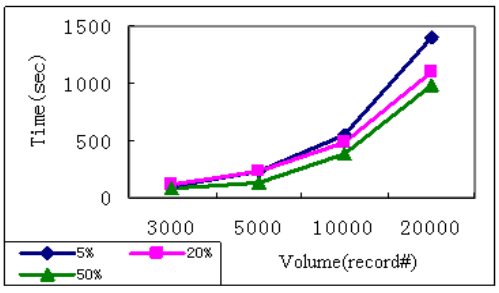


Figure 1. Performance test for multiple data-set & supports

V. Conclusion

DCIP provides a new and efficient algorithm for discovering MFI; it condenses data-set by deleting items in infrequent 1-itemsets and merging duplicate transactions repeatedly, and utilizes the intersections of (1-s)*|D|+1 transactions with other transaction itemsets to perform pruning; along with the discovering process, with the increasing of the number of deleted transactions, the number of times needed for calculating intersections will decrease rapidly. It's time & space cost increases not drastically when data-set volume increases, so its usability retains for MFI applications for high volume data-sets.

The DCIP algorithm can be further optimized in various aspects, such as keep a record of all resulting intersections to avoid duplicated generation of identical intersections to further improve the efficiency of this algorithm.

References

[1] Ceglar A, Roddick JF. "Association Mining". *ACM Computing Surveys*, 2006, 38 (2), pp.1–42

[2] Rigoutsos L, Floratos A. "Combinatorial Pattern Discovery in Bio-logical Sequences: The Teiresias Algorithm". *Bioinformatics*, 1998, 14 (1):,pp.55–67

[3] Bayardo RJ. "Efficiently Mining Long Patterns from Databases". In: Haas LM, Tiwary A, eds. *Proceedings ACM SIGMOD International Conference on Management of Data*, 1998, pp.85–93

[4] Lin DI, Kedem ZM. "Pincer-Search: A New Algorithm for Discovering the Maximum Frequent Set". In: Schek HJ, et al., eds. *Proceedings of 6th International Conference on Extending Database Technology*, 1998, pp.105–119

[5] Agarwal RC, Aggarwal CC, Prasad VVV. "Depth First Generation of Long Patterns". In: Ramakrishnan R, Stolfo S, eds. *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, pp.108–118

[6] Burdick D, Calimlim M, Gehrke J. "MAFIA: A Maximal Frequent Itemset Algorithm for Transactional Databases". In: Georgakopoulos D et al, eds. *Proceedings of the 17th International Conference on Data Engineering*, 2001, pp.443–452

[7] Gouda K, Zaki MJ. "Efficiently Mining Maximal Frequent Itemsets". In: Cercone N, Lin TY, Wu XD, eds. Proceedings of the 2001 IEEE International Conference on Data Mining, 2001, pp.163–170

[8] Song Yu-qing, Zhu Yu-quan, Sun Zhi-hui, etc. "An Algorithm and Its Updating Algorithm Based on FP-Tree for Mining Maximum Frequent Itemsets". *Journal of Software*, 2003, 14(9), pp.1586-1592. (in Chinese)

[9] Li Qing-hua, Wang Hui, Jiang Cheng-yi. "Parallel Algorithm for Mining Maximal Frequent Itemsets", *Computer Science*, 2004, 31 (12), pp. 132–134,188 (in Chinese)

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 016-019

Research Survey on Integrated Software Engineering Environment Based on Product Line

Jianli Dong
School of Computer Engineering , HuaiHai Institute of Technology , Lianyungang, 222005,China.
dongjl1019@sina.com

***Abstract*—Through research and analysis on the software engineer process and the life cycle model based on product line, integrated software engineering environment model, core resource and the environment database platform, the technology of automation of modern manufacturing industry production line and so on, a novel multi-layers and multi-dimensional integrated software engineering environment model and realizing framework based on product line is proposed , this model takes the product line core assets components as the middleware and bus. At the same time, with the unified conceptual model, data model and behavior model, the integrated mechanism of the environment interfaces, tools and data in the model is systematically studied and discussed. The models, theories, technologies and methods mentioned in this paper have significant reference and guidance effect on the research and development of current product line software engineering and integrated software engineering environment.**

***Index Terms*—integrated software engineering environment, software components, software architecture, software product line, core assets, environment database platform**

I. INTRODUCTION

In resent years, with the application and the growing maturity of new technologies such as the software architecture, software component, large granularity software reuse, the engineering and automation production methods of the application software products' mass customization in specific domains based on the integration of these technologies, that is product line software engineering methodology, has been developed and researched. At the same time, it has been a matter of social concern and high priority in software engineering and industry and become the focus of the study in the current field of software engineering. The engineering approaches based on product line will lead to a dramatic change of the software development methods, software development, from the traditional "algorithms + data structure + manual coding" one-time production methods that begin with zero, will step across into the automated production mode which has modern manufacturing industries characteristics of "software architecture + software components + assembly line". The achievement of software product line approach must rely on the product line engineering development environment with the characteristics of the product line and production capacity, that is ISEE-BPL : Integrated Software Engineering Environment Based on Product Line. So it is the achievement of software engineering methodology and integrated software engineering environment based on product line, which is very important significance to the development of modern software industry. And it had become the important move to promote information industry and maintain a sustained and rapid growth of the social economy for the world nations.[1-4].

II. THE CORE TECHNOLOGIES ON ISEE-BPL RESEARCH

It has essential differences between integrated software development environment based on product line and the common software development environment created by structure-based or object-oriented software engineering approach, the traditional development environment for short. Integrated software engineering environment based on product line, achieve mass customization of software products and automation production in specific areas, according to the mode of production "software architecture + component + assemble", while traditional development environment is a kind of one-time development methods starting with zero in the light of the mode of production "data structure + algorithms+ manual coding", and the practical extent, productivity and restructuring update of the environment would receive a great deal of restrictions for its commonality. As modern manufacturing industry cannot use a same production line to produce difference productions in different fields such as saloon car, plane and television, it is greatly at odds with the development direction and production methods of modern industry. So integrated software engineering environment based on product line would be the most ideal software product line to realize the industrialization and automation of software production nowadays.

To research and realize integrated software engineering environment based on product line, the basic structure and mechanisms of integrated environment should be studied and cleared. It mainly include research on software engineering process model based on product line, research on integrated software engineering environment model based on product line, research on core resources data model and database platform based on product line, and research on the realizing mechanism and key technology of integrated software engineering environment on basis of product line. Then to realize the environment of interface, tools and the natural integration of data on the basis of creating unified conceptual model, data model

This project is supported by Fund of Jiangsu University Natural Science Basic Research Project, Grant No. 08KJD520013.

and behavioral model. What we mentioned above is the concern of this paper.

A. Software Engineering Process and Life-Cycle Model Based on Product Line

Essentially, software engineering environment based on product line is a kind of product line which similar to the automatic production line of modern manufacturing industry. It is also a new software engineering method and process to carry out mass customization production of software products in specific domain based on standard component of core resources such as software architecture, component, connecting piece, production plan, specification, constraint, documents and so on. Therefore, what the most important for research on the product line software development environment is to set up software development process model and life-cycle model which suitable for the characteristics and the production methods of product line. It is used to describe the whole process of products development based on product line systematically, and then take this as a guide to determine the message-based application, tool configuration and production process.

Its goal is to describe the sequence of activities, workflow, the task framework, product submission and standards of software engineering process based on product line completely, clearly and specifically. And the guidelines to action and behavioral norms to implement the software product line engineering and software products would be the prerequisite and an important foundation for the research on the integrated software development environment. In recent years, there have been some preliminary research results on the research of product line engineering process model. For example: software product line double life cycle model and SEI model[2]. But there simple models can hardly meet the requirement of the whole process expressing ability of modern software management system, mode of producing, evolution of e-Learning, quality control and so on, such as the Multi-level upper and lower layer organization and management system of international, national, industry, domains and application and so on which owned by product line project, the engineering process characteristics and mode of multi-level iterative production methods and the evolution of multi-dimensional product.[5-6].

On the research and creation of the product line engineering process models and life-cycle, we firstly propose a kind of opened "N-life cycle model" suitable for software engineer based on product line. This model contain the whole process of product line software engineer, each operational phase division of inter process ,the customization of task framework, product quality standards, the entire process of monitoring the completion steps, management and technical characteristics completely. Compared with product line double life cycle model, SEI model and so on, N-life-cycle model, an open process model, which use for reference of the modern industry process and management system and has been proved to be more features and manage spatial of modern industry, meet the product line software project process modeling and expression ability[7].

B. The Model of ISEE-BPL

It is almost blank for the real product of integrated software engineering environment based on product line, which is the source language of American CMU/SEI expert. But the study of integrated environment model is a quite active research direction in the software engineering field with the development of software technology. The research emphasis is to construct the architecture, framework and model of integrated environment, so as to its integration mechanism and implementations. At present, the sanctioned integrated reference models is 3-dimensional model based on network distributed computing environment, which is equal to interface, tools and data integration. But limited in the traditional software engineering methodology, it is common software development environment and disposable and zero starting point software product develop. and this environment model cannot get enough to the "architecture + component +component assembly" software development methods and production capacity based on product line, which is also the main basis why the CMU/SEI evaluation product line development environment is almost blank[8-9]. Except having the low level and source code development ability of traditional software engineering environment, integrated software engineering environment model based on product line should also provide the basic characteristics of "territoriality, abstractness, publicity, expansibility, variability and reusability" possessed by core resource such as architecture and component, as so as system-level components automated assembly capacity possessed by product family mass customized production. This is the essential difference between product line software development environment and traditional software development environment, and it is also the key issues and primary target that must be solved and achieved by integrated software engineering environment based on product line.

In the process of research and established integrated software engineering environment model based on product line, the product line software engineering environment model, which has the open property and layered architecture, must be designed at the basis of unified conceptual product line engineering model, unified large granularity reusable core resource data model, unified component composition production behavior model, the unified routing model of mutually iterative of core resource development and software products production. At the light of layered architecture, the environment models can be divided into three part of interface layer, tool layer and data layer in sequence. Interface layer, which is used to receive user information and request and implement the tool transferring and the return of result data, is to carry out interface integration and management; tool layer, which is used to provide the service for interface layer and realize data access and

sharing, is to carry out tool integration and management; data layer in fact is environment database platform, which is used to realize data integration, storage and management of environment data. Here, it is important to point out that architecture layer based on product line integrated development environment model implies the two part of software production environment based on product line and software development in traditional, that is, take the product line component as bus or middleware, its above is product line software engineering environment to realize high level and system level software products assembly and automatic production, while its blew is traditional software development environment to realize low level and source code level product line component programming and development. It is the essential difference of product line software development environment and traditional software development environment, and the important thought and method innovation of this paper [10].

C. *Product Line Core Assets Database System and Environment Database Platform*

Software product line is a kind of new software engineering method and software development paradigm formed by domain engineering, software architecture, software component and software reuse technology. Software product line mainly consisting of two part of core assets and application products, while core assets includes large complex heterogeneous of domain-specific software architecture, component, connector, production planning, development document, test plan, use case, standard specification and constraint and so on and reusable software production resources. So an important basic research of integrated software engineering environment based on product line is product line core assets and the designed and realized of integrated environment database platform. Its main research contents includes: the study of the data model possess of the ability of core resource data expression and description, product line core assets database schema design, the research of operation mechanism and management ability on product line core assets storage, classification, retrieval, query, version, reuse and optimization and so on. That is, in the demand of capacity, except having the conventional database storage and management capabilities, core assets database should provide rich and reused production resource and convenient resource retrieval, query, reuse, assembly and configuration and so on and management ability for software products customization, assembly and production, this is also the basic goal of the research on product line core assets and environment database system[11-12].

Currently, the most common and extensive database model is relational data model in the design of product line complex heterogeneous core resource database model, but for the capability defect of relation model such as poor expression ability and semantic fault, the establish of product line core assets database model must adopt object-oriented model that having ultra-intense expression ability and good mechanism, to complete the expression of complex heterogeneous resource data and the demand of modeling; in the design of core assets database, research and establish the multi-view core resource database mode which either has the features of three layer structure of product line architecture style, architecture framework and architecture component or mapping view, reuse degree view and relational view, to realize product line engineering characters and management requirements database, here not only provide management function of core assets such as storage, classification, retrieval, version, optimization and configuration, it also provide good mechanism and methods of data integration, tool integration and interface integration. Furthermore, to consider current network running environment, core assets database platform should offer Internet proxy function to realize localization and network of component integration. These are the basic characteristics and ability that environment database system must have.

D. *Framework and Realization Mechanism of ISEEBPL*

It is key to research and establish multi-dimension integrated environment model and multi-level environment architecture in the design and realization of product line integrated development environment overall framework. Multi-dimension is that software production should support the iterative, evolution and constraint of product line management project, enterprise product line project, domain engineering, application engineering and different stage of each engineering process in the two dimensions of vertically and horizontally. Multi-level is macroscopic level division of integrated environment, that is, take the software product line core assets component as the bus and middleware, its above is product line software engineering environment and should support high-level and software production based on system-level component assembly, for below, it could support the design and development of low-level and manual encoding method based source code program. All these indicate that the software development environment that based on product line must have the automated and industrialization productivity to carry out high-level and system-level component assembly technology, at the same time, it has the ability of low-level and source code grade manual codes development component resource program, which is its multi level and the essential difference between product line and traditional software development environment, which is called double environment integration.

Integrated environment should develop and distribute the management and development tools that satisfying the needs of the task in each stage in the light of the model regulations and requirement. and then implementing the layer-by-layer integration of interface, tools and data according to integrated environment model. Interface layer implement the integration and management of interface, its role is to receive users information and request and to carry out tools call and user feedback and submission of result data or products according to the production process of software product line. The tool layer is to implement the integration and management of

tools, it offer the service for interface layer and realize the access and sharing of data for next layer.

The tools consists of the management tools based on product line project(including industrial organization management, standard compliance management, enterprise management, production process management, core assets and product configuration management and so on), domain analysis modeling tool, domain architecture design tool, domain component development and reengineering generation tool, software products assembly tool, core resource database management and maintenance tool, system structure test and analysis tool and so on. Not only do these tools carry out tool integration according to integrated environment, but also the organic integration of tool layer and data layer, tool layer and interface layer. By ensuring the tools call, scheduling, communicate, collaboration and interoperate, tool integration should guarantee the access and sharing of data for tools, as so as offering complete service for users. The realization mechanism and technology research on integrated software engineering environment based on product line mainly realize the natural integration of environment interface, tools and data according to establishing unified "conceptual model, data model and behavior model". Especially, the tool plug and play integration and the realization of database platform and environment data integration including product line core resources would be the key technology and emphasis issue that need to solve in the research and development of integrated software engineering environment.[13-15]

III. CONCLUSION

To sum up, basis of advance software theory, technology and method in resent years, and taking the research and realization of integrated software engineering environment based on product line as the theme, and taking the realization of industrial and automated production of software products as the goal, this paper gives a systematic summarization of the research contents and key technology of product line engineering model, integrated environment model, core assets and environment database platform, the framework and realization mechanism of environment and so on that mentioned with product line integrated software engineering environment, which represents the research focus and the frontier of modern software engineering methodology, play an essential foundation role and guiding significance in the forming of modern software engineering and the development of modern software industry. Of course, we should clearly see that the research on integrated software engineering environment based on product line still has a long way to go, compared with the automatic production line of modern manufacture. So we still need to continue and make a great effort, to make the initiative work for realizing the industrial and automatic production of software products.

REFERENCE

[1] Yang Fu-Qing, “Thinking on the Development of Software Engineering Technology”, Journal of software(in Chinese), 2005,16(1), pp.1-7.

[2] Paol. Clements, Linda Nort-hrop(America, the Original Author), Zhang Li, Wang Lei(China, Translators), “Software Product Lines: Practices and Patterns”, Tsinghua University Press(Beijing), 2003.

[3] Sun Chang-ai, Jin Mao-zhong, Liu Chao, “Overviews on Software Architecture Research”, Journal of Software(in Chinese), 2002,Vol.13(7), pp.1228-10.

[4] Wang Zhijian, “Software component technology and application(in Chinese)”, Science Press(BeiJing), 2005.

[5] Samuel A. Ajila, Ali B. Kaba, “Evolution support mechanisms for software product line process”, Journal of Systems and Software, Vol81(10), pp.1784-1801,October 2008.

[6] Daniel Mellado, Eduardo Fernández-Medina, Mario Piattini, “Towards security requirements management for software product lines: A security domain requirements engineering process”, Computer Standards & Interfaces, Vol.30(6), pp.361-371, 2008.

[7] Dong Jian-li, “Research on software engineering process model based on software product line architecture”, Computer Engineering and Design(in Chinese), 2008,29(12):3016-3018.

[8] Jintae Kim, Sooyong Park, Vijayan Sugumaran, “A framework for domain requirements analysis and modeling architectures in software product lines”, Journal of Systems and Software, Vol.81(1), January 2008, pp.37-55.

[9] P. Lempp, “Integrated computer support in the software engineering environment EPOS — Possibilities of support in system development projects”, Microprocessing and Microprogramming, Volume 18, Issues 1-5, December 1986, pp.223-232.

[10] Jianli Dong, Jianzhou Wang, “The Research of Software Product Line Engineering Process and It’s Integrated Development Environment Model”, ISCSCT-2008 Proceeding, IEEE Computer Society, Vol.1, pp.66-71.

[11] S Chen, J. M. Drake, W. T. Tsai, “Database requirements for a software engineering environment: criteria and empirical evaluation. Information and Software Technology”, Vol.35(3), March 1993, pp.149-161.

[12] Paul Brown, “Distributed component database management systems”, Component Database Systems, 2001, pp.29-70.

[13] IEEE, Inc. “Information Technology--Guideline for the Evaluation and Selection of CASE Tools (IEEE Std 1462-1988). New York, NY:IEEE Computer Society Press, 1998.

[14] Hitchcock, P Inf, “Introduction to integrated project support environments”, Software Technol. Vol.29(1), 1987, pp.15–20.

[15] Minder Chen, Ronald J. Norman, “A framework for integrated CASE”, IEEE Software, No.3, 1992, pp.18-22

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 020-025

SimHash-based Effective and Efficient Detecting of Near-Duplicate Short Messages

Bingfeng Pi, Shunkai Fu, Weilei Wang , and Song Han
Roboo Inc., Suzhou, P.R.China
{winter.pi, shunkai.fu, willer.wang, song.han}@roboo.com

***Abstract*—Detecting near-duplicates within huge repository of short message is known as a challenge due to its short length, frequent happenings of typo when typing on mobile phone, flexibility and diversity nature of Chinese language, and the target we prefer, near-duplicate. In this paper, we discuss the real problem met in real application, and try to look for a suitable technique to solve this problem. We start with the discussion of the seriousness of near-duplicate existing in short messages. Then, we review how SimHash works, and its possible merits for finding near-duplicates. Finally, we demonstrate a series of findings, including the problem itself and the benefits brought by SimHash-based approach, based on experiments with 500 thousands of real short messages crawled from Internet. The discussion here is believed a valuable reference for both researchers and applicants.**

***Index Terms*—Near-duplicate, SimHash, short text**

I. INTRODUCTION

Duplicate and near-duplicate web documents are posing large problems on Web search engines: They increase the space required to store the index, slow down serving results, and annoy the users [2, 3]. Among the data available on Internet, a large proportion are short texts, such as mobile phone short messages, instant messages, chat log, BBS titles etc [1]. It was reported by Information Industry Ministry of China that more than 1.56 billion mobile phone short messages are sent each day in Mainland China [5]. Being an active and popular mobile search service provider in China, our history query log indicates that the short message search enjoys a similar scale of monthly PVs (Page Visit) as Web page search on Roboo® [6]. These two vivid facts motivate us to pay enough attention to the quality of our short message repository since it is the basis for quality search service. Unfortunately, the status of duplicate or near-duplicate messages is very severe, especially near-duplicates. For example, the following are two typical examples near-duplicates (all in Chinese):

- In the first pair, the one above has 4 more characters (highlighted in gray) than the other one, and the remaining part is exactly the same;
- And in the second pair, the differences contain one character, and two punctuations (all highlighted in gray).

These differences may result from several causes: 1) same contents appearing on different sites are all crawled, processed and indexed; 2) mistake introduced while parsing these loosely structured and noisy text (HTML page may contain ads., and it is known as shorting of semantics useful for parsing); 3) manual typos (all information on Internet are created by people originally) and manual revising while being referred and reused; 4) explicit modification to make the short message suitable for difference usage (for example, replacing "春节" (Spring Festival, Chinese traditional New Year) with "新年" (Near Year), though they are actually similar in meaning.

Manual checking may be applicable when the scale of repository is small, e.g. hundreds or thousands of instances. When the amount of instances increases to millions and more, obviously, it becomes impossible for human beings to check them one by one, which is tedious, costly and prone to error. Resorting to computers for such kind of repeatable job is desired, of which the core is an algorithm to measure the difference between any pair of short messages, including duplicated and near-duplicated ones.

Manku et al. [3] showed that Charikar's SimHash [4] is practically useful identifying near-duplicates in web documents. SimHash is a fingerprint technique enjoying the property that fingerprints of near-duplicates differ only in a small number of bit positions. A SimHash fingerprint is generated for each object. If the fingerprints of two objects are similar, then they are deemed to be near-duplicates. As for a SimHash fingerprint f, Manku et al. developed a technique for identifying whether an existing fingerprint f' differs from f in at most k bits. Their experiments show that for a repository of 8 billion pages, 64-bit SimHash fingerprints and $k = 3$ are reasonable. Another work by Pi et al. [2] confirmed the effect of SimHash and the work by Manku et al; besides, they proposed to do the detection among the results retrieved by a query, i.e. so-called query-biased approach. It reduces the problem scale via divide-and-conquer, replacing global search with local search, and it is open to more settings possibly met in application, e.g. smaller k to remove fewer documents under some condition, and bigger k to delete more documents under other condition.

In this paper, we show that SimHash is indeed effective and efficient in detecting both duplicate (with $k = 0$) and near-duplicate (with $k > 0$) (see the two typical examples in TABLE II.) among large short message repository. However, we also notice that due to the born feature of short messages, $k = 3$ may not be an ideal parameter for. For example, as shown in TABLE III. , $k = 2$ is enough to detect the one-character

difference, but k has to be 5 to detect the same pair of messages with two-character difference. Besides, with the same one-character difference, short messages require larger k for effective detection (TABLE IV.). This may be explained by an observation, that the same difference, e.g. having one different character on the same position of two short messages, would be more influential to short text than to long text.

This is a paper focusing on discussing practical solution for real application, and our contribution is three-fold. Firstly, we demonstrate a series of practical values of SimHash-based approach by experiments and our experience. Secondly, we point out that $k = 3$ may be suitable for near-duplicated Web page detection, but obviously not suitable for short messages. Thirdly, we propose one empirical choice, $k = 5$, as applied on our online short message search (http://wap.roboo.com). In Section 2, we describe how SimHash works, its advantages and disadvantages. Then in Section 3, we present a series of experiments, and discuss the results. A brief review of conventional work is presented in Section 4, followed by conclusion and future work in Section 5.

TABLE I. TYPICAL NEAR-DUPLICATES OF SHORT MESSAGES, WITH DIFFERENCES HIGHLIGHTED IN GRAY

(1) **春节搞笑春节搞笑**祝福短信新年到了，事儿多了吧？招待客人别累着，狼吞虎咽别撑着，啤的白的别掺着，孩子别忘照顾着，最后我的惦念常带着。新年快快乐乐的！！ (2) **春节搞笑**祝福短信新年到了，事儿多了吧？招待客人别累着，狼吞虎咽别撑着，啤的白的别掺着，孩子别忘照顾着，最后我的惦念常带着。新年快快乐乐的！！
(1) 又是你的生日了，虽然残破的爱情让我彼此变得陌生，然而我从未忘你的生日，happy birthday (2) 又是你的生日了，虽然残破的爱情让我们彼此变得陌生，然而我从未忘你的生日。Happy birthday!

TABLE II. EXAMPLE: DETECT DUPLICATE WITH $k = 0$ AND NEAR-DUPLICATE WITH $k > 0$ (WITH DIFFERENCES HIGHLIGHTED IN GRAY)

$k = 0$	(1) 今生今世，你是我唯一的选择。愿我们好好珍惜缘分，也请你答应我，今生今世只为我守候。 (2) 今生今世，你是我唯一的选择。愿我们好好珍惜缘分，也请你答应我，今生今世只为我守候。
$k > 0$	(1) 今生今世，你是我唯一的选择。愿我们好好珍惜缘分，也请你答应我，今生今世只为我守**候**。 (2) 今生今世，你是我唯一的选择。愿我们好好珍惜缘分，也请你答应我，今生今世只为我守**侯**。

TABLE III. EXAMPLE: DETECT SAME LONG TEXT BUT MORE DIFFRENCE REQUIRES LARGER k (WITH DIFFERENCES HIGHLIGHTED IN GRAY)

$k = 2$	(1) 今生今世，你是我唯一的选择。愿我们好好珍惜缘分，也请你答应我，今生今世只为我守**候**。 (2) 今生今世，你是我唯一的选择。愿我们好好珍惜缘分，也请你答应我，今生今世只为我守**侯**。
$k = 5$	(1) 今生今世，你是我唯一的选择。愿我们好好珍惜缘分，也请你答应我，今生今世只为我**守候**。 (2) 今生今世，你是我唯一的选择。愿我们好好珍惜缘分，也请你答应我，今生今世只为我**等待**。

TABLE IV. EXAMPLE: DETECT SAME DIFFRENCE BUT SHORTER TEXT REQUIRES LARGER k (WITH DIFFERENCES HIGHLIGHTED IN GRAY)

$k = 2$	(1) 今生今世，你是我唯一的选择。愿我们好好珍惜缘分，也请你答应我，今生今世只为我守**候**。 (2) 今生今世，你是我唯一的选择。愿我们好好珍惜缘分，也请你答应我，今生今世只为我守**侯**。
$k = 5$	(1)愿我们好好珍惜缘分，也请你答应我，今生今世只为我守**候**。 (2)愿我们好好珍惜缘分，也请你答应我，今生今世只为我守**侯**。

I. NEAR-DUPLICATE DETECTION BY SIMHASH

A. SimHash and Hamming Distance

Charikar's SimHash [4], actually, is a fingerprinting technique that produces a compact sketch of the objects being studied, no matter documents discussed here or images. So, it allows for various processing, once applied to original data sets, to be done on the compact sketches, a much smaller and well formatted (fixed length) space. With documents, SimHash works as follows: a Web document is converted into a set of features, each feature tagged with its weight. Then, we transform such a high-dimensional vector into an $f-bit$ fingerprint where f is quite small compared with the original dimensionality. An excellent comparison of SimHash and the traditional Broder's shingle-based fingerprints [7] can be found in Henzinger [8].

To make the document self contained, here, we give the algorithm's specification in Figure 1. , and explain it with a little more detail. We assume the input, document D , is pre-processed and composed with a series of features (tokens). Firstly, we initialize an f -dimensional vector V with each dimension as zero (line 1). Then, for each feature, it is hashed into an $f-bit$ hash value. These f bits increment or decrement the f components of the vector by the weight of that features based on the value of each bit of the hash value calculated (line 4-8). Finally, the signs of the components determine the corresponding bits of the final fingerprint (line 9-11).

```
SimHash( document D)
{
01  Init vector Sim[0..(f-1)] = 0;
02  For (each feature F in document D) Do
03      F is hashed into an f-bit hash value X;
04      For (i = 0; i < f; i++) Do
05          If (X[i] == 1) Then
06              Sim[i] = Sim[i] + weight(F);
07          Else
08              Sim[i] = Sim[i] - weight(F);
09  For (i = 0; i < f; i++) Do
10      If (Sim[i] > 0) Then Sim[i] = 1;
11      Else Sim[i] = 0;
}
```

Figure 1. Algorithm specification of SimHash.

SimHash has two important but somewhat conflicting properties: (1) The fingerprint of a document is a "hash" of its features, and (2) Similar documents have similar hash values. The latter property is quite different from traditional hash function, like MD5 or SHA-1 (Secure Hash Algorithm), where the hash-values of two documents may be quite different even they are slightly different. This property makes SimHash an ideal technique for detecting near-duplicate ones, determining two documents are similar if their corresponding hash-values are close to each other. The closer they are, the more similar are these two documents; when the two hash-values are completely same, we actually find two exact duplicates, as what MD5 can achieve.

In this project, we choose to construct a $64-bit$ fingerprint for each web document because it also works well as shown in [1]. Then the detection of near-duplicate documents becomes the search of hash values with $k-bit$ difference, which is also known as searching for nearest neighbors in hamming space [3, 4]. How to realize this goal efficiently? One solution is to directly compare each pair of SimHash codes, and its complexity is $O(N^2)$, where N is the size of document repository and each unit comparison needs to compare 64 bits here. A more efficient method as proposed in [1] is implemented as well in this project. It is composed of two steps. Firstly, all $f-bit$ SimHash codes are divided into $(k+1)$ block(s), and those codes with one same block, say 1,2, …, $(k+1)$, are grouped into different list. For example, with $k=3$, all the SimHash codes with the same 1st, 2nd, 3rd, or 4th block are clustered together. Secondly, given one SimHash code, we can get its 1st block code easily and use it to retrieve a list of which all codes sharing the same 1st block as the given one. Normally, the length of such list is much smaller than the whole size of repository, N. Besides, given the found list, we need only check whether the remaining blocks of the codes differ with k or fewer bits. The same checking need to be applied to the other 3 lists before we find all SimHash codes, i.e. all near-duplicate documents. This search procedure is referred as hamming distance measure by us in the remaining text.

B. Advantages and Disadvantages of SimHash

SimHash has several advantages for application based on our experience:

1. Transforming into a standard fingerprint makes it applicable for different media content, no matter text, video or audio;
2. Fingerprinting provides compact representation, which not only reduces the storage space greatly but allows for quicker comparison and search;
3. Similar content has similar SimHash code, which permits easier distance function to be determined for application;
4. It is applicable for both duplicate and near-duplicate detection, with $k=0$ and $k>0$ respectively;
5. Similar processing time for different setting of k if via the proposed divide-and-search mentioned above, and this is valuable for practice since we are able to detect more near-duplicates with no extra cost;
6. The search procedure of similar encoded objects is easily to be implemented in distributed environment based on our implementation experience;
7. From the point of software engineering view, this procedure may be implemented into standard module and be re-used on similar applications, except that the applicants may determine the related parameters themselves.

Standard and aligned encoded output (e.g., 64-bit SimHash code) plus the parameter k make it possible to figure out flexible, re-usable and scalable near-duplicate detecting algorithm, like the one implemented in [1,2] and this project as well. TABLE II. , TABLE III. and TABLE IV. demonstrate several near-duplicated pairs detected with SimHash. The difference of each pair of short messages and the corresponding k value required for the detection are listed. As we discussed here, SimHash can be applied to short text without any modification on our previous work on page document, i.e. long text. Besides, it is noticed that $k=0$ lets us to find exact duplicates, and larger k allows us to detect more difference.

However, SimHash has its weak points as well. The text length has great influence on the effect. For example, $k=2$ allows us to find the pairs with one different character in TABLE III. , but it requires $k=5$ in TABLE IV. . Of course, we are lucky to cost similar computing time with different k, but we have to tradeoff manually on the choice of k since determining whether or not near-duplicated is quite vague especially for those detected with large k. Besides, the size of the target objects being studied has influence on the choice of k. That's why we can't directly apply $k=3$ here though it is proved effective in our Web page cleaning project.

II. EXPERIMENTAL STUDY

This is a project aiming at discussing practical solution for real-world large scale application, so it is believed that experiments with real data are highly desired. In this section, we are going to cover the following aspects:

- The algorithm is effective to find both duplicates and near-duplicates among short-text repository;
- The problem of near-duplicate is serious, so it is worthy of our effort;
- $k=3$ is not good choice for detecting near-duplicated short texts;
- SimHash-based approach is flexible, customizable and scalable.

A. Our Data

We crawl and parse Web pages, extracting and indexing about 500 thousands of short messages for experimental study. Too short messages are filtered first, and the minimum threshold value is 20 here. Note that

this choice is arbitrary. TABLE V. summarizes the testing repository.

TABLE V. BASIC STATISTICS ABOUT THE SHORT MESSAGE REPOSITORY USED FOR TESTING

# of messages with length of ≥ 20	498,959
Length of longest message	2,968
Length of shortest message	20
Mean length	85.64
Standard deviation	182.44

B. Correctness and Effectiveness

To make the following discussion sound, it is necessary to verify the algorithm and our implementation. From the examples shown in TABLE II. , we notice that:

- Duplicate pairs are indeed detected with $k = 0$;
- Near-duplicated pairs have to be detected with $k > 0$;
- Larger difference requires larger k;
- If one near-duplicate can be detected will smaller k, definitely it can be detected by larger k. The reverse is not true. Therefore, with $k > 0$, we can find both duplicate and near-duplicates.

Other than these sample examples, we further randomly select 1000 messages from the whole repository. With $k = 3$, 65 near-duplicated pairs are found, and they are checked one by one manually. The conclusion is that all 65 pairs are indeed near-duplicates.

C. Seriousness of Near-duplicate Problem

To reflect the seriousness of both duplicate and near-duplicate among short message repository, we conduct the search with different k on the same test repository respectively. Figure 2 shows the number of near-duplicated pairs detected given different k, ranging from 0 to 10. It is noticed that there are 87,604 pairs detected with $k = 0$, i.e. duplicate message. It means about 35% (87604 * 2 / 498959) of total messages are exactly duplicated (note: our hashing is token-based, and space is ignored.) This rate increases to about 57% when $k = 10$, that is more than half are duplicated or near-duplicated. With such many near-duplicates existing in the repository, we can image the quality of search result – a series of same or similar results are piled together and presented to the users. Because normally there is no extra score，like PageRank score in page search, but the similarity score to consider given short message search application, we have no way to improve the user experience but filtering out those duplicated ones.

Figure 2. also confirms the discussion in Section 3.2, i.e. larger k allows us find more near-duplicates. By removing those duplicated and near-duplicated ones, storage space is reduced greatly as well. Besides, by reducing the index scale, the online retrieval response should be quick. Therefore, there are several benefits if we are able to delete those repeating texts.

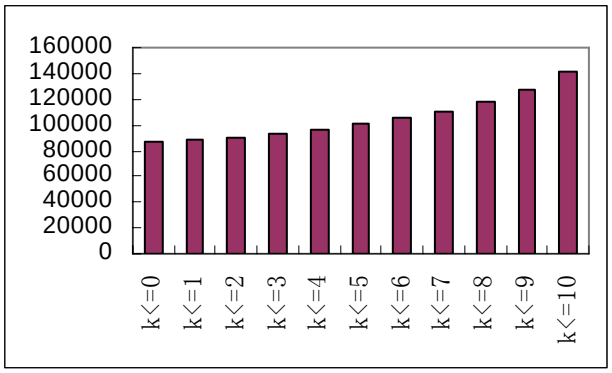


Figure 2. The total number of near-duplicated pairs detected with different k, ranging from 0 to 10.

D. $k = 3$ is not Good Choice

$k = 3$ is demonstrated in [1] and [2] as suitable and practical choice for large scale near-duplicate detection of Web document. However, it seems not appropriate for detecting near-duplicated short messages. As the examples of TABLE III. and TABLE IV. indicate, we may only detect one-character difference even with $k = 5$. In other words, although the pair of short messages is quite similar one another, differ in only one character, they are not detected as near-duplicated with $k = 3$. Still with the experiments shown in Figure 2, the ratio of near-duplicates detected to all is about 37% for $k = 3$. That means that about extra 20% (57% - 37%, and 57% is the ratio when $k = 10$) near-duplicates are left in the short message repository if we apply $k = 3$.

Why the same $k = 3$ doesn't work well enough on short text? It can be explained in a not so formal way. Given a Web page with 1000 characters, and a short text with 50 characters, the influence of adding or deleting or changing one character will have much less influence on the Web page than on the short text. This is also the feature of fingerprinting technique, and we can improve the sensitivity by using more fits while constructing the fingerprint. However, it is not free lunch since the corresponding computing and storage burden is increased meanwhile.

Given 64-bit fingerprint, which k is most appropriate is hard to decide in practice. Though we may do a similar experiments like in [1], asking some persons to check manually the number of true positives, true negatives, false positives and false negatives, it is not employed here due to three causes:

- It is very costly a procedure, in term of money and time;
- Whether or not near-duplicate is not easy to determined in many cases, even by human beings, say nothing of machine;
- It is not our and any similar service providers' goal to remove "all" near-duplicates. A practical goal is to alleviate the influence on users to an affordable level.

On our online short-message search service (http://wap.roboo.com), $k = 5$ is taken, and the general evaluation by several month-person testing is satisfactory, much better than before when there is no any action is taken on the message repository. Figure 3. is the snapshot of our online search of short message, and the result list appearing on the right screen is clean, no duplicated or near-duplicated. This is meaningful since (1) the small screen is made full use of by only displaying unique results; (2) it saves the communication flow for users by displaying no repeating ones; (3) the user is able to find what s/he like in a quicker manner (fewer times of paging down).

Figure 3. The home page of our short-message search (left, accessible via http://wap.roboo.com), and the result list given query “春节”(Spring Festival, Right).

E. SimHash-based Approach is Flexible, Customizable and Scalable

From the discussion above, we can see that SimHash-based near-duplicate detection algorithm allows us to find both duplicate and near-duplicate ones, which owes to its most nature – similar object has similar SimHash code. It is not only applicable to Web documents, but short messages here. The only necessary adjustment is to find a suitable k. Applicants may customize the choice of k based on their goals, i.e. how strictly we want to control the result. Of course, increasing the value of k also increases the risk of removing false negatives.

Our experiments with about 500 thousands of data are done on a common PC machine, with 3.06GHz CPU and 1GB memory. Each experiment only cost us dozens of seconds to do the search, and the time is similar for different k. In real production environment, we implement a Hadoop-based [9] version, which allows us easily scale to millions of cases with few machines. Actually, we notice that the encoding, grouping and the comparison procedure is easy to be programmed with MapReduce [10] framework, one famous divide-and-conquer distributed computing model.

III. RELATED WORK

A variety of techniques have been proposed to identify academic plagiarism [11, 12, 13], Web page duplicates [2,3,8, 14] and duplicate database records [15,16]. However, it is noticed that there are very few works on the discussion of detecting near-duplicates among short text repository until recently, including [1, 17]. Gong et al. [1] proposed the SimFinder which employ three techniques, namely, the ad hoc term weighting, the discriminate-term selection and the optimization techniques. It is a fingerprinting-based method as well, but takes some special processing while choosing features and their corresponding weights. Muthmann et al. [17] discussed the near-duplicate detection for Web forums which is another critical resource of user-generated content (UGC) on Internet. It is also built on the basis of fingerprinting technique. However, there is no article about the related work on mobile search application upon preparing this paper.

Though the theoretical basis may be similar, identification of near-duplicate short messages is believed much more difficult considering that: 1) it usually contains less than 200 characters, and there are few effective features to extract; 2) it tends to be informal and error prone; 3) the degree of duplicated and near-duplicated is known as more severe than Web documents. All these can be explained by the fact that short messages are very popular and welcome by mobile users, and they are so short to be distributed easily.

IV. CONCLUSION AND FUTURE WORK

While providing short message search service, we are short of other reference, like the measures by PageRank®, to optimize the ranking of results retrieved, but their relative similarity to the query itself. Based on traditional search model, same or similar short messages may pile together in the result list. Besides, it is noticed and near-duplicates are abundant in short text database. Both facts together motive us to pay enough attention to detecting and eliminating them, to ensure the user experience. We review SimHash, and discuss the application of SimHash in detecting near-duplicated short text. SimHash has several advantages, and we prove them based on a series of experiments with real data.

Deleting both duplicated and near-duplicate contents has several benefits, especially, for mobile application like us, including that (1) allow more useful information to present on the small screen; (2) save the time and bandwidth for users by reducing the possible times of paging down operation or asking the server for a new page; (3) reduce the storage requirement; (4) reduce the online retrieval time, so as the waiting time of users. User experience will never be over-emphasized on mobile application considering the small screen, difficult inputting and slow connection speed today. It is believed that our discussion here may be valuable reference for applicants like us since our own product is benefiting from this technique currently online.

Although there is no special operation taken to process the features in our system, like those appearing in [1,17], it is observed that the existing framework works well online. However, we also notice that there is space there for improvement. For example, we may for further to study the relationship of text length, ratio of difference and suitable k ’s option. Besides, some advanced NLP

(Natural Language Processing) techniques may be applied to improve the outcome. For instance, we may recognize and fix typo first before applying SimHash encoding, which is possible to allow us to find more difference with same k . Of course, all extra finer modeling will be paid with more computing resource.

REFERENCES

[1] C. Gong., Y. Huang., X. Cheng. and S. Bai., "Detecting Near-Duplicates in Large-Scale Short Text Databases," Proc. of PAKDD 2008, LNAI, vol. 5012, pp. 877-883. Springer, Heidelberg

[2] B. Pi., S.-K. Fu., G. Zou, J. Guo. and H. Song, "Query-biased Near-Duplicate Detection: Effective, Efficient and Customizable," Proc. of 4th International Conference on Data Mining (DMIN), Las Vegas, US., 2008.

[3] G.S. Manku, A. Jain. and A.D. Sarma, "Detecting Near-Duplicates for Web Crawling," Proc. of 16th International World Wide Web Conference (WWW), 2007.

[4] M. Charikar, "Similarity Estimation Techniques from Rounding Algorithm," Proc. of 34th Annual Symposium on Theory of Computing (STOC), 2008, pp 380-388.

[5] Official website of Chinese Information Industry Ministry of China: http://www.mii.gov.cn/.

[6] Roboo mobile search engine: http://wap.roboo.com/.

[7] A.Broder, , S.C. Glassman, M. Manasse. and G. Zweig, "Syntactic clustering of the web," Computer Networks, vol.29, no.8-13, 1997, pp 1157-1166.

[8] M.R. Henzinger, "Finding near-duplicate web documents: a large-scale evaluation of algorithms," Proc. of ACM SIGIR, 2006, pp 284-291.

[9] Hadoop official site: http://hadoop.apache.org/core/.

[10] J. Dean and S.Ghemawat, "MapReduce: Simplified data processing on large cluster," Proc. of 6th Symposium on Operating System Design and Implementation (OSDI), 2004.

[11] S.Brin, J.Davis and H.Garcia-Molina, "Copy detection mechanisms for digital documents," Proc. of the ACM SIGMOD Annual Conference, San Francisco, CA, 1995.

[12] N.Shivakumar and H.Garcia-Molina, "SCAM : A copy detection mechanism for digital documents, " Proc. of 2nd International Conference in Theory and Practice of Digital Libraries, Austin, Texas, 1995.

[13] M.Zini, M.Fabbri and M.Mongelia. "Plagiarism detection through multilevel text comparison," Proc. of the 2nd International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution, Leeds, U.K., 2006.

[14] N.Shivakumar and H.Garnia-Molina, "Finding near-replicas of documents on the web," Proc. of Workshop on Web Databases, Valencia, Spain, 1998.

[15] Z.P.Tian, H.J.Lu and W.Y.Ji, "An n-gram-based approach for detecting approximately duplicate data records," International Journal on Digital Libraries, 5(3):325-331, 2001.

[16] M.A. Hernandez and S.J.Stolfo, "The merge/purge problem for large databases," Proc. of ACM SIGMOD Annual Conference, San Jose, CA., 1995.

[17] K.Muthmann, W.M.Barczynski, F.Brauer and A.Loser,"Near-duplicate detection for web-forums,", 142-152, International Database Engineering and Applications Symposium(IDEAS), 2009.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 026-029

The Matrix Reduct Algorithm of Incomplete Decision Table

Wenjun Liu,and Zhuo Long
Department of Mathematics and Computing Science, Changsha
University of Science and Technology, Changsha, China
liuwjzhlp@126.com , lz860108@163.com

***Abstract*—First, we give the definition of relative matrix of fuzzy relative, then, Combining the technology of fuzzy clustering and ideal of rough sets, a method to obtain the attribute reduct from an incomplete decision table which includes continuous, order and discrete attributes is put forward.. Example shows this algorithm is effective and feasible.**

***Index terms*—continuous domain decision table, fuzzy sets, ,rough sets, attribute reduct, matrix.**

I. INTRODUCTION

Rough set theory was proposed by Z. Pawlak in 1982 [1]. It is a new mathematical tool to deal with imprecise, incomplete and inconsistent data. After more than twenty years of pursuing rough set theory and its application, the theory has reached a certain degree of maturity. In recent years we have witnessed a rapid growth of interest in rough set theory and its application worldwide [2-3].

In rough set theory and its application, attribute reduct of decision table is an important research field. Many experts all over the world are pursuing this research. And there are lots of successful examples of attribute reduct[4-8]. On the base of paper[8], in this paper, combining clustering technic and rough set theory, we put forward an attribute reduct algorithm of incomplete continuous decision table .

II. PRELIMINARIES

In rough set theory, knowledge is look as the classification ability of objects. Suppose we are given a finite set $U \neq \phi$ of objects we are interested in. Suppose R is an equivalence relation over U, then by U/R we mean the family of all equivalence classes of R, and $[x]_R$ denotes an equivalence class of R containing an element $x \in U$. With each subset $X \subseteq U$, we associate two subsets:

$$\underline{R}X = \bigcup\{Y \in U/R \mid Y \subseteq X\},$$
$$\overline{R}X = \bigcup\{Y \in U/R \mid Y \cap X \neq \phi\}$$

called the R-lower and R-upper approximations of X respectively.

Let $T = (U, A, C, D)$ （in short T） be a decision table, With condition attributes C and decision attributes D, we can define the C positive region $\mathrm{pos}_C(D)$ in the equivalence relation $\mathrm{ind}(D)$ as: $\mathrm{pos}_C(D) = \bigcup_{X \in U/ind(D)} \underline{C}X$.

An attribute $a \in C$ is called D-dispensable, if $\mathrm{pos}_C(D) = \mathrm{pos}_{C-\{a\}}(D)$, else a is called D-indispensable. The significance of attribute $a \in C$ is defined as:

$$\delta^D_{C-\{a\}}(a) = \frac{|\mathrm{pos}_C(D) - \mathrm{pos}_{C-\{a\}}(D)|}{|U|}.$$

If every attribute of C is D-indispensable, then C is called independent of D. The subset $B \subseteq C$ is called a D- reduct of C if B is independent of D and $\mathrm{pos}_B(D) = \mathrm{pos}_C(D)$.

.A decision table T is called consistent, if $\mathrm{pos}_C(D) = \boldsymbol{U}$, else it is called inconsistent. The consistent degree of decision table T is defined as: $\gamma = \frac{|\mathrm{pos}_C(D)|}{|U|}$. Obviously, a decision table is consistent if and only if its consistent degree $\gamma = 1$.

Definition 1[8] Let R be an equivalence relation over U, $|U| = n$, the relative matrix of R is defined as $M_R = (r_{ij})_{n\times n}$, where $r_{ij} = \begin{cases} 1 & x_i R x_j \\ 0 & \text{else} \end{cases}$.

Obviously, M_R is a symmetrical matrix, and all the elements of main diagonal are 1.

Definition 2[8] Let $M_1 = (r_{ij})_{n\times n}$ and $M_2 = (r_{ij}')_{n\times n}$ be two relative matrixes, then $M_1 \cap M_2$ is defined as: $M_1 \cap M_2 = (s_{ij})_{n\times n}$, $s_{ij} = \min\{r_{ij}, r_{ij}'\}$.

Definition 3 Let $M_1 = (r_{ij})_{n\times n}$ and $M_2 = (r_{ij}')_{n\times n}$ be two relative matrixes, if for every $i, j \in \{1,2,\cdots,n\}$, $r_{ij} \le r_{ij}'$, then we called $M_1 \le M_2$; else $M_1 \not\le M_2$.

Obviously, decision table T is a consistent decision table if and only if $M_C \le M_D$, where $M_C = \bigcap_{a\in C} M_a$,

$M_D = \bigcap_{d \in D} M_d$.

In decision table T , $|U| = n$, let $M_C = (r_{ij})_{n \times n}$ and $M_D = (r_{ij}')_{n \times n}$ be two relation matrixes of condition attributes C and decision attributes D relatively, denotes $M_{EX} = \bigcup \{i, j \mid r_{ij} > r_{ij}', i < j\}$.

Definition 4 In decision table T , the dependency degree of decision attributes D with respect to condition attributes C is defined as: $\gamma_C(D) = 1 - \frac{|M_{EX}|}{|U|}$.

Obviously, decision table is a consistent decision table if and only if $\gamma_C(D) = 1$.

Definition 5 Let T be a decision table, the significance of condition attributes $C' \subseteq C$ with respect to decision attributes D is defined as: $\sigma_D(C') = \gamma_C(D) - \gamma_{C-C'}(D)$. Especially, if $C' = \{a\}$, the significance of condition attribute a with respect to decision attributes D is: $\sigma_D(a) = \gamma_C(D) - \gamma_{C-\{a\}}(D)$.

If $\sigma_D(a) \neq 0$, then attribute a is D -indispensable; else a is D -dispensable.

Theorem 1 For a consistent decision table T , a condition attribute $a \in C$ is D -dispensable if and only if: $M_{C-\{a\}} \leq M_D$.

Proof We can easily know from the definitions of relation matrix and dispensable attribute.

Definition 6 The subset of condition attributes $C' \subseteq C$ is called a D - reduct of C , if:

i) $\gamma_{C'}(D) = \gamma_C(D)$;

ii) for every $a \in C'$, $\sigma_D(a) \neq 0$.

According to the definition of D - reduct and the structure of relative matrix, we can easily know:

Theorem 2 The subset of condition attributes $C' \subseteq C$ is a D -reduct of C , if:

i) $M_{C'} = M_C$;

ii) for every $a \in C'$, $M_{C'-\{a\}} \not\leq M_C$.

III. The Attribute Reduct Algorithm of Incomplete Decision Table

Definition 7 Decision table T is called an incomplete continuous domain decision table, if there exists a continuous condition attribute a_i, and there at least exists $a_j \in C$, such that V_{a_j} have null values(where V_{a_j} is the attribute values set of objects on attribute a_j), but decision attributes D are all discrete attributes and for each $d \in D$, V_d have no null value. In here, we used "*" express null value.

Definition 8 Let T be an incomplete continuous domain decision table, $V_a = \{x_1, x_2, \cdots, x_s\}$, if a is a discrete attribute, then the similarity degree of two attribute values x_i and x_j $(i, j \in \{1,2,\cdots,s\})$ with respect to attribute a is defined as:

$$\mu_a(x_i, x_j) = \begin{cases} 1 & x_i = x_j \vee x_i = "*" \vee x_j = "*" \\ 0 & \text{else} \end{cases};$$ if a is an ordinal attribute, then the similarity degree of two attribute values x_i and x_j with respect to attribute a is defined as:

$$\mu_a(x_i, x_j) = \begin{cases} 1 - \frac{|x_i - x_j|}{s-1} & x_i, x_j \neq "*" \\ 1 & x_i = "*" \vee x_j = "*" \end{cases},$$ where s is the number of ordinal attribute values; if a is a continuous attribute, then the similarity degree of two attribute values x_i and x_j with respect to attribute a is defined as:

$$\mu_a(x_i, x_j) = \begin{cases} 1 - \frac{|x_i - x_j|}{a_{\max} - a_{\min}} & x_i, x_j \neq "*" \\ 1 & x_i = "*" \vee x_j = "*" \end{cases},$$ where $a_{\max}$ and $a_{\min}$ are the maximum and minimum values of attribute a respectively.

Definition 9 Let T be an incomplete continuous domain decision table, $|U| = n$, $C = \{a_1, a_2, \cdots, a_m\}$ are condition attributes, D are decision attributes. The similarity degree of two samples u_i and u_j is defined as: $r_{ij} = \min_{l=1}^{m} \mu_{a_l}(x_{il}, x_{jl})$, where x_{il}, x_{jl} are the attribute values of objects u_i and u_j on attribute a_l .

In an incomplete continuous domain decision table, obviously, the similarity degree r_{ij} of objects u_i and u_j is in [0,1], that is the similarity degree matrix $R = (r_{ij})_{n \times n}$ is a fuzzy similarity matrix.

Definition 10 Let $R = (r_{ij})_{n \times n}$, for every $\lambda \in [0,1]$, $R_\lambda = (r_{ij}(\lambda))_{n \times n}$ is called the λ -cut matrix of matrix R , where $r_{ij}(\lambda) = \begin{cases} 1 & r_{ij} \geq \lambda \\ 0 & r_{ij} < \lambda \end{cases}$.

We know, rough set theory has a greatest merit that it does not need any additional information about data. In the following, combining rough set theory and fuzzy

clustering technic, we put forward an attribute reduct algorithm of incomplete continuous domain decision table.

If T is an incomplete continuous domain decision table, $C=\{a_1,a_2,\ \cdots,a_m\}$ are condition attributes, which include continuous, discrete and ordinal attributes, $D=\{d\}$ is a discrete decision attribute, the processes of attribute reduct algorithm are as such:

(1) Compute M_d according to definition 1;

(2) Compute the similarity degree of each object with all condition attributes according to definition 9 and build up the fuzzy similarity matrix R;

(3) Turn R into an equivalence matrix $\hat{R}$ [9,10], then ascertain the threshold of classification according to $\hat{R}$, the method of ascertaining the threshold is as such:

i) sorted the values in $\hat{R}$ from big to small, denotes $\alpha_1>\alpha_2>\cdots>\alpha_k$ (k is the number of different values in $\hat{R}$; ii) let $i=1$; iii) for $\lambda_i\in(\alpha_{i+1},\alpha_i]$, compute $\hat{R}_{\lambda_i}$, if $\hat{R}_{\lambda_i}\not\leq M_d$, then turn to v); iv) if $i<k-1$, then $i\leftarrow i+1$, goto iii); v) if $i=1$ the threshold of fuzzy clustering is α_1; else, the threshold of fuzzy clustering is α_{i-1}.

(4) Initialize the attribute reduct set: let $B=C$;

(5) Let $j=1$;

(6) For every $a_j\in B$, compute the fuzzy similarity relation R_{a_j} of objects about attributes $B-\{a_j\}$, turn the fuzzy similarity relation into an equivalence relation $\hat{R}_{a_j}$ and obtain its α_{i-1} -cut matrix $(\hat{R}_{a_j})_{\alpha_{i-1}}$, according to definition 5, compute $\gamma_{B-\{a_j\}}(D)$. If $\gamma_{B-\{a_j\}}(D)=\gamma_B(D)$, then $B=B-\{a_j\}$; if $\gamma_{B-\{a_j\}}(D)\neq\gamma_B(D)$, then the significance of a_j is: $\sigma_D(a_j)=\gamma_B(D)-\gamma_{B-\{a_j\}}(D)$;

7) If $j<n$, then $j=j+1$, goto 6);

8) Put out the attribute reduct B.

Obviously, under the same threshold, if removing an attribute will change the positive, it means that the attribute is significant and indispensable, else it is dispensable.

V. EXAMPLE

Table 1 is an incomplete continuous domain decision table, a_1,a_2 are continuous condition attributes, a_3 is a discrete condition attribute. $D=\{d\}$ is a discrete decision attribute. In the following, we compute the attribute reduct used the algorithm we put forward above.

TABLE 1.
AN INCOMPLETE CONTINUOUS DOMAIN DECISION TABLE

U	a_1	a_2	a_3	d
1	0.9	2	1	1
2	1.1	0.8	1	0
3	1.3	3	2	0
4	1.4	1	1	1
5	1.4	2	1	1
6	1.2	1	1	1
7	1.8	3	2	0
8	4	3	2	0
9	*	3	2	0
10	1.3	*	1	1
11	2	1	*	0

Firstly, build up the fuzzy similarity matrix R_C (the matrix built up by the similarity degree of each object with all condition attributes) and M_d . Through computing, we have the following R_C and M_d.

THE FUZZY SIMILARITY MATRIX R_C

$$\begin{pmatrix}
1 & 0.45 & 0 & 0.54 & 0.84 & 0.54 & 0 & 0 & 0 & 0.88 & 0.54\\
 & 1 & 0 & 0.9 & 0.45 & 0.91 & 0 & 0 & 0 & 0.94 & 0.71\\
 & & 1 & 0 & 0 & 0 & 0.84 & 0.13 & 1 & 0 & 0.09\\
 & & & 1 & 0.54 & 0.94 & 0 & 0 & 0 & 0.97 & 0.54\\
 & & & & 1 & 0.54 & 0 & 0 & 0 & 0.97 & 0.54\\
 & & & & & 1 & 0 & 0 & 0 & 0.97 & 0.74\\
 & & & & & & 1 & 0.29 & 1 & 0 & 0.09\\
 & & & & & & & 1 & 1 & 0 & 0.09\\
 & & & & & & & & 1 & 0 & 0.09\\
 & & & & & & & & & 1 & 0.77\\
 & & & & & & & & & & 1
\end{pmatrix}$$

MATRIX M_d

$$\begin{pmatrix}
1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0\\
 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1\\
 & & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1\\
 & & & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0\\
 & & & & 1 & 1 & 0 & 0 & 0 & 1 & 0\\
 & & & & & 1 & 0 & 0 & 0 & 1 & 0\\
 & & & & & & 1 & 1 & 1 & 0 & 1\\
 & & & & & & & 1 & 1 & 0 & 1\\
 & & & & & & & & 1 & 0 & 1\\
 & & & & & & & & & 1 & 0\\
 & & & & & & & & & & 1
\end{pmatrix}$$

Secondly, ascertain the classification threshold.

i) compute the transitive closure of R_C, we obtain the fuzzy equivalence matrix $\hat{R}_C$ as such:

$$\begin{pmatrix} 1 & 0.88 & 0.09 & 0.88 & 0.88 & 0.88 & 0.09 & 0.09 & 0.09 & 0.88 & 0.81 \\ & 1 & 0.09 & 0.94 & 0.94 & 0.94 & 0.09 & 0.09 & 0.09 & 0.94 & 0.81 \\ & & 1 & 0.09 & 0.09 & 0.09 & 1 & 1 & 1 & 0.09 & 0.09 \\ & & & 1 & 0.97 & 0.97 & 0.09 & 0.09 & 0.09 & 0.97 & 0.81 \\ & & & & 1 & 0.97 & 0.09 & 0.09 & 0.09 & 0.97 & 0.81 \\ & & & & & 1 & 0.09 & 0.09 & 0.09 & 0.97 & 0.81 \\ & & & & & & 1 & 1 & 1 & 0.09 & 0.09 \\ & & & & & & & 1 & 1 & 0.09 & 0.09 \\ & & & & & & & & 1 & 0.09 & 0.09 \\ & & & & & & & & & 1 & 0.81 \\ & & & & & & & & & & 1 \end{pmatrix}$$

ii) According to the definition of λ -cut matrix, we can get $(\hat{R}_C)_1 \le M_d$, $(\hat{R}_C)_{0.97} \le M_d$, but $(\hat{R}_C)_{0.94} \not\le M_d$. According to the method of ascertaining threshold, we can get the classification threshold is 0.97.

Thirdly, let $B = C$. Delete condition attribute a_1 from decision table, under the threshold λ =0.97, through computing, we can get $(\hat{R}_{B-\{a_1\}})_{0.97} \not\le M_d$, that is a_1 is D -indispensable, and we obtain the significance of a_1 is:

$\sigma_D(a_1) = \gamma_B(D) - \gamma_{B-\{a_1\}}(D)$ =0.57.

In the same way, we can get: under the threshold λ =0.97, a_2 and a_3 are both D -indispensable, the significance of a_2 is 0.45; the significance of a_3 is 0.73.

At last, obtain the attribute reduct. According to the above algorithm, we can get the attribute reduct of this decision table is $\{a_1, a_2, a_3\}$.

Since for a decision table with continuous attributes, it hardly appears inconsistent case, so in step 3, we request that the consistent degree $\gamma = 1$ is reasonable. In order to add to the ability of fault tolerance, we can give a consistent degree less than 1, such as 98%, the algorithm is similarly.

In the continuous domain decision table, in order to diminish the affect come from different physical measure,we can normalize the continuous attribute values before we compute similarity matrix.

V. Conclusion

In rough set theory, attribute reduct is an important research field. For a given discrete decision system, at present, there are many valid algorithms of acquiring attribute reduct. For an incomplete decision system with continuous, ordinal and discrete attributes, before computing the significance and attribute reduct of it, we always complete and discretize it.However, the information will be lost during these process. In order to decrease the information lose, in this paper, we put forward a method to obtain attribute reduct directly from incomplete decision table with continuous, ordinal and discrete attributes. And also, we verified the validity of thisattribute reduct algorithm through an example.

Acknowledgment

This work is supported by Natural Science Foundation of China (10701018). The authors are grateful for the reviewers who made constructive comments.

References

[1] Pawlak Z., Rough Set, International Journal of Computer and Information sciences , 1982, 11(5), pp.341--356.

[2] Pawlak Z., Reasoning about data-A rough set perspective, LNAI 1424, Springer,Bolan, 6, 1998.

[3] Skowron A.,Rough sets in KDD, Special Invited Speaking, WCC 2000 in Beijing, Aug, 2000.

[4] Wang G. Y., Rough set theory and knowledge acquire, Xi'an JiaoTong University Press, Xi'an,2001.

[5] Zhang W. X., Wu W. Z. and Liang J. Y., Rough set theory and method, Science Press, Beijing, 2001.

[6] Miao D. Q., Wang J., Rough sets based approach for multivariate decision tree construction, Journal of Software 1997(6),pp.425--431.

[7] Liang J. Y., The algorithm on knowledge reduction incomplete information systems international journal of uncertainty, Fuzziness and Knowledge-Based Systems, 2002(10), pp. 95--103.

[8] Guan J. W., Bell D. A. and Guan Z., Matrix computation for information systems}, Information Sciences, 2001(3),pp. 129--156.

[9] Liu P. Y., Wu M. D., Fuzzy theory and its application, The National University of Defense Technology Press, 1998.

[10] Yang L. B., Gao Y. Y., Fuzzy mathematic principle and application , South China University of Technology Press, Guangzhou, 1997

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 030-034

Efficiently Methods for Embedded Frequent Subtree Mining on Biological Data

Wei Liu[1,2] , Ling Chen[2] , and Lan Zheng[1]
[1]School of Mathematics Information Technology, Nanjing Xiaozhuang University , Nanjing, China
[2]Department of Computer Science, Yangzhou University, Yangzhou, China
Email: yzliuwei@126.com, lchen@yzcn.net

***Abstract*—As a technology based on database, statistics and AI, data mining provides biological research a useful information analyzing tool. The key factors which influence the performance of biological data mining approaches are the large-scale of biological data and the high similarities among patterns mined. In this paper, we present an efficient algorithm named IRTM for mining frequent subtrees embedded in biological data. We also advanced a string encoding method for representing the trees, and a scope-list for extending all substrings for frequency test. The IRTM algorithm adopts vertically mining approach, and uses some pruning techniques to further reduce the computational time and space cost. Experimental results show that IRTM algorithm can achieve significantly performance improvement over previous works.**

***Index Terms*—Embedded Frequent Sub Tree; Scope -List; Biological data;**

I. INTRODUCTION

Mining frequent patterns, including mining transactions, sequences, trees and graphs, is a fundamental and important problem in data mining area[1,2]. As one increases the most complexity of the structures to be discovered, one extracts more informative patterns in bioinformatics; researchers are especially interested in mining tree patterns. Mining tree patterns is widely used in the areas such as bioinformatics, web-mining, chemical data structure mining etc. For example, in bioinformatics, researchers have collected vast amounts of RNA structures, which are essentially trees. To obtain information about a newly sequenced RNA, they compared it with known RNA structures[3], looking for common topological patterns, which provide important clues to the function of the RNA.

In recent years, extensive efforts have been devoted to developing efficient algorithms for frequent subtree pattern mining. But most, if not all, of the proposed algorithms are Apriori-based[4-7] algorithms with breadth-first search. Nevertheless, another kind of methods for mining subtrees adopted depth-first search strategy. In[8], Zaki proposed two more efficient algorithms for embedded frequent subtree discovery in an ordered forest: TreeMiner and PatternMatcher. For the same problem, Wang[9] developed two efficient pattern-growth methods: Chopper and Xspanner algorithms. Is it straightforward to apply these methods into biological data mining? Unfortunately, since the biological data to be handled are mainly about DNA sequences or protein sequences with primary structure, but all frequent pattern mining methods used are limited to some easy frequent item mining or frequent sequential mining algorithms[10] , they are not suitable for more complicated structures.

In this paper, based on the above idea, we present an efficient algorithm named IRTM for mining frequent subtrees embedded on biological data. The IRTM algorithm adopts vertically mining approach, and uses some pruning techniques to further reduce the computational time and space cost. The experimental results show that IRTM algorithm is more efficient and scalable in comparison to the traditional TreeMiner and PatternMatch algorithms.

Identify applicable sponsors: Chinese National Natural Science Foundation under grant No. 60673060, Natural Science Foundation of Jiangsu Province under contract BK2008206 and Foundation of Nanjing Xiaozhuang University under grant No.2008NXY44.

II. PROBLEM DEFINITION AND CONCEPTS

A. The problem of mining frequent subtrees

Node numbers and labels A tree is an acyclic connected graph denoted as $T(v_0, N, L, B)$, where N is the set of labeled nodes in T, $v_0 \in N$ is the root node of T, L is the set of the labels of the nodes, B is the set of directed edges in T. We define the size of T as the number of its nodes, which is denoted as $|T|$. Each node in N has a well-defined index, i, according to its position in a depth-first(or pre-order) traversal of the tree. We use the notation $n(v_i)$ to refer to the ith node according to the indexing scheme(i=0…$|T|$-1). The label (also referred to as an *item*) of each node is taken from a set of labels L={0,1,2,3,…,m-1} where m is the number of different labels. For a node $v_i \in N$, its label is denoted as $L(v_i)$ whereas its number is denoted as $n(v_i)$. Each edge $b=(v_x, v_y) \in B$ is an ordered pair of nodes, where v_x is the parent of v_y . Defferent nodes in a tree may carry an identical label.

Hereafter, without special mention, the iterm "tree" refers to the labeled, ordered, and rooted ones.

Mining Frequent subtrees Given a database D consisting of n trees and a user specified minisuport *minsup*, mining frequent subtrees is to efficiently enumerate all frequent subtrees.

B. Concepts and definitions

Node scope Assuming that a tree T has g nodes. For a node $v_i (0 \le i \le g-1)$ in T, we use T_i to denote the subtree with v_i as its root. By a preorder traversal scheme, starting

from node v_i and ttraversing the subtree T_i, we can get a node sequence. Let the last node in the sequence be v_r, then v_r must be the rightmost leaf nodes in subtree T_i. We call $s(v_i)=[n(v_i),n(v_r)]$ the scope of node v_i. Figure 2-1 shows an example of a tree and the numbers, labels and scopes of its nodes.

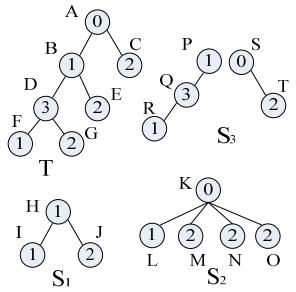


Figure 2-1 Tree and subtree

L(A)=0 s(A)=[0, 6]
L(B)=1 s(B)=[1, 5]
L(D)=3 s(D)=[2, 4]
L(F)=1 s(F)=[3, 3]
L(G)=2 s(G)=[4, 4]
L(E)=2 s(E)=[5, 5]
L(C)=2 s(C)=[6, 6]

Figure 2-2 labels and scopes of nodes

In Figure 2-1, we can see the set of labels in T is L={0,1,2,3} and the set of its nodes is $N=\{A,B,C,D,E,F,G\}$. According to its position in a depth-first(or pre-order) traversal of T, the number of each node is shown as follows:

Node	A	B	C	D	E	F	G
Number	0	1	6	2	5	3	4

And the label of each node, and its scope are shown in Figure 2-2.

String encoding method Let $\Psi(T)$ be the node sequence generated by traversing the tree T in the preorder. The string encoding of tree T, which is denoted by $C(T)$, can be produced as follows: initially $C(T)$ is an empty set, then we create the node sequence $\Psi(T)$ by traverse the T in the preorder scheme. Whenever we add a node v_i into $\Psi(T)$ during the traversing, the opening form of its label $\overline{L(v_i)}$ is added into $C(T)$. When we trace back from node v_i to its parents, the closing form of its label $\underline{L(v_i)}$ is also added into $C(T)$. Since each node should be traversed twice, the length of $C(T)$ is $2|T|$. Since each tree T has a unique string encoding $C(T)$, and each string encoding C represents only one tree structure, this string encoding method can completely preserve the structural information of a tree.

As shown in Figure2-1, S_1 is an embedded subtree of T. To generate the string encoding of $S_1$by the above method, the string encoding of S_1 is $C(S_1)$={ $\bar{1}\ \bar{1}\ \underline{1}\ \bar{2}\ \underline{2}\ \underline{1}$ } and $C(S_2)$={ $\bar{0}\ \bar{1}\ \underline{1}\ \bar{2}\ \underline{2}\ \bar{2}\ \underline{2}\ \bar{2}\ \underline{2}\ \underline{0}$ }. Because S_3 is unconnected, it is not a subtree but a sub-forest.

C. *Substring validity checking*

In our algorithm, when a new candidate string C' is generated, because only extendable substrings are extended in the algorithm, we should verify whether C' is extendable, i.e., the validity of C'.

A valid substring must satisfy the following two conditions:

(1) For each $\bar{i}$ in the substring, it must be matched by an exclusive $\underline{i}$ in the substring.

(2) If there is an $\bar{i}$ before $\bar{j}$ in the substring, their corresponding matching characters $\underline{i}$ and $\underline{j}$ must satisfy the condition: $\underline{i}$ appears after $\underline{j}$ in the substring .

For instance, the string encoding of tree T in Figure 2-1 is $C(T)$= { $\bar{0}\ \bar{1}\ \bar{3}\ \bar{1}\ \underline{1}\ \bar{2}\ \underline{2}\ \underline{3}\ \bar{2}\ \underline{2}\ \underline{1}\ \bar{2}\ \underline{2}\ \underline{0}$ } , its substrings { $\bar{1}\ \bar{1}\ \underline{1}\ \bar{2}\ \underline{2}\ \underline{1}$ } and { $\bar{0}\ \bar{1}\ \underline{1}\ \bar{2}\ \underline{2}\ \bar{2}\ \underline{2}\ \bar{2}\ \underline{2}\ \underline{0}$ } are valid substrings since they represent subtrees of T.

Given a tree database D, we can build a guide tree to store all subtree information for the trees in it. And then we can mine the common subtree in D by the guide tree. We can find all substrings by gradually expansion on the string encoding of guide tree from left to right and thereby can find all subtrees of the originally tree database.

Guide Tree Let $D=\{T_1, T_2, \ldots, T_n\}$ denote a database consists of n trees, its guide tree T is an rooted tree which satisfies the following conditions:

(1) The node set of T is absolutely same to the node set of all frequent subtrees in D;

(2)For any two nodes, if and only if they are ancestor-descendant relationship in at least one tree of D and there also exists the same relationship between the two nodes in T;

(3)There is no common path started from the root with length more than 0;

For example, as shown in figure 2-3, A is a guide tree but B is not.

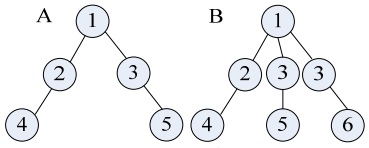


Figure 2-3 Guide Tree

III. ALGORITHMS

A. *The frame of our algorithm IRTM*

In this section we present an algorithm named IRTM (Intensive Rooted Tree Mining) to handle the problem of biological data mining. The algorithm first builds a guide tree according to the input tree database, and constructs the string encoding for all of the trees using the string encoding method. Then, the algorithm uses the guide tree for efficient candidate subtrees generation. The algorithm expands a candidate subtree T_i by a rightmost expansion on its string encoding $C(T_i)$. In such rightmost expansion on $C(T_i)$, we simply attach a character on the end of $C(T_i)$ to form a new string C'. When a new candidate string C' is generated, we should verify if it is expandable. We'll build scope-lists for all expandable substrings when they are generated and frequency tested. Repeat the whole process until all the frequent subtrees are discovered. By the combination of extension, scope-list building and frequency testing, we can greatly reduce the search space and improve performance of the algorithm.

In the above algorithm, during the extension process, we only simply add nodes with different labels on the rightmost of the string encoding. But the candidate subtrees produced by the method will often not exist in

any subtree of *D* and thus it will cost large computing. Then we use guide tree to greatly reduce the number of candidate subtrees.

Let the roots of all trees in *D* have the same label. (while not, we can add a subnode labeled'*' on the roots of all trees) Firstly for each tree, we should prune all infrequent nodes and rebuild a new tree. During pruning process, the original ancestor-descendant relationship among other nodes should be retained.

For example, assuming that the node *X* will be deleted from the tree, we can directly regard its child node(u,v,w) as *Y*'s child.

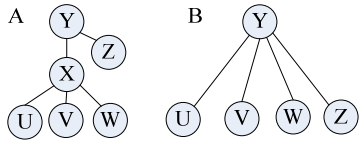


Figure 2-4 node pruning and tree rebuilding operation

As shown in figure 2-4(A), while deleting node *X*, the tree will be like figure2-4 (B). If there is no father for *X*, we'll split the tree into several subtrees. Each of them will be rooted as one child of *X*.

After pruning all infrequent nodes of *D*, we can build guide tree by the following method:

1. Let $T = \varnothing$
2. For each tree T_i of *D* do
3. For each leaf *l* of tree T_i do
4. Let the path from the root to *l* be *r*
5. if $T = \varnothing$ then add *r* into *T*
6. else
7. In *T*, we can find a path *s* with the common prefix of *r* among all paths from the root to all leaves. Let *r*=(*a*, *g*) and *s*=(*a*, *h*). The vertex *x* is the last vertex of *a*, then *g* can be regarded as one sub-path started from *x* and be added into *T*.
8. End if
9. End for
10. End for

The framework of our algorithm IRTM is shown as follows:

Input：a tree database *TDB*
a support threshold *min_sup*
the set *P* of all substrings

Output：the set *T* of all embedded frequent subtrees

Begin

1.Scan database TDB once, generate its guide tree *T*;
2.Generate the string encoding *C*(*T*)of *T* and
let $C(T) = \{x_1, x_2, ...x_n\}$
3. Build the initial Scope-list for each vertex;
4. $P = \{\varnothing\}$;
6. For *i*=1 to *n* do
7. For each string *Y* in *P* do
8. Combine *Y* with x_i to a new string $y' = \overline{Yx_i}$;
9. Use string extension checking for *y'* ;
10. If *y'* is an extendable substring Then
11. Build the scope-list of *y'* by using scope-lists of *y* and x_i
12. If x_i is $\bar{s}$ Then
13. Construct a subtree *t'* corresponding with *y'*
14. Frequency test for *t'*
15. If *t* 'is frequent Then add *y'* into *P* and add *t* ' into*T* ;
16. Else
17. Add *y'* into *P*
18. End if
19. End if
20. End for
21. End for

End

B. The production of Scope-List

In line 3 and 11 of the algorithm IRTM, scope-list of a newly generated string *y'* is constructed. Here we give the definition of the scope-list as follows：

Let *y* represents a subtree *T*(*y*) with *k* nodes, each entry in its scope-list can be denoted as $[t,s]$, where *t* stands for the tree where *T*(*y*) is embedded, and $s=[l_1,u_1][l_2,u_2]...[l_k,u_k]$ denotes the scopes of the nodes in the tree.

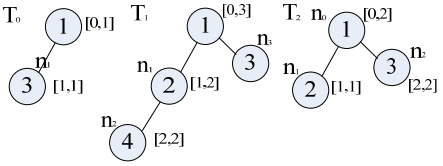


Figure 3-1 A forest with three subtrees of T

Using the trees shown in Figure 3-1 as the data set, we illustrate the procedure of constructing the scope-lists (*S-L* in short) in the algorithm. Since the algorithm generates the candidate set from the nodes of the trees, it first builds the initial scope-lists for the nodes with different labels as follows:

$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$
0[0,1] 1[0,3] 2[0,2]	1[1,2] 2[1,1]	0[1,1] 1[3,3] 2[2,2]	1[2,2]

To expand the candidate subtrees, then we should detect the relations between the scopes of the subtree and the node expanded.

Assuming that the scopes of a subtree *x* and node *y* in a tree are $s_x = [l_x, u_x]$ and $s_y = [l_y, u_y]$ respectively, if and only if $l_x \le l_y$ and $u_x \ge u_y$, we call s_y is a subscope of s_x, which is denoted as $s_x \supset s_y$.

In algorithm IRTM, when a candidate string *C* is extended by attaching a character *g* to generate a new string $\overline{Cg}$, the *S-L* of $\overline{Cg}$ can be constructed as follows:

(1) Find all the entry pairs with the same *t* value in S-L's of *C* and *g*;

(2) Assuming $(t,[l_{11},u_{11}], [l_{12},u_{12}], \ldots, [l_{1k},u_{1k}])$ and $(t,[l_2,u_2])$ is such a pair and $[l_{1k},u_{1k}] \supset [l_2,u_2]$, then an entry $(t,[l_{11},u_{11}], [l_{12},u_{12}], \ldots, [l_{1k},u_{1k}], [l_1,u_1])$ should be added into the *S-L* of $\overline{Cg}$.

From the rules above we can see that for each entry $(t,[l_{11},u_{11}], [l_{12},u_{12}], \ldots, [l_{1k},u_{1k}])$ in the *S-L* of C, there must be $[l_{11},u_{11}] \subset [l_{12},u_{12}] \subset \ldots \subset [l_{1k},u_{1k}]$.

In the data set in Figure 3-1, let the candidate string be $C=\{\bar{1}\}$ and the character attached be $g=\bar{2}$.By the above rules, we can get the newly constructed *S-L* of $\{\bar{1}\ \bar{2}\}$ as follows:

$\bar{1}$	$\bar{2}$
1 [0,3] [1,2] 2 [0,2] [1,1]	

Similarly , for another instance, the newly constructed *S-L* of $\{\bar{1}\ \bar{2}\ \bar{4}\}$is as follows:

$\bar{1}$	$\bar{2}$	$\bar{4}$
1 [0,3] [1,2] [2,2]		

When a candidate string *C* is expanded by attaching a character $\underline{x}$, we can get the *S-L* of the new string simply by deleting the last scope in the entry on *S-L* of *C*. For instance, in the data set in Figure 3-1, let the candidate string be $C=\{\bar{1}\ \bar{2}\ \bar{4}\}$ and the character attached be $\underline{4}$.Then the scope-list of $\{\bar{1}\ \bar{2}\ \bar{4}\ \underline{4}\}$ is as follows :

$\bar{1}$	$\bar{2}$	$\bar{4}$	$\underline{4}$
1 [0,3] [1,2]			

C. Frequency Test

Given a tree database *TDB*, a subtree *T'* and its corresponding string *P'*, the support of *T'* is the number of trees in *TDB* that contain at least one occurrence of *T'*. According to the definition of the scope-list mentioned above, an entry $(t,[l_{11},u_{11}],\ \ldots,\ [l_{1k},u_{1k}])$ in the *S-L* of *P'* indicates that *T'* appears once in the tree *T* of TDB. And we can easily get the support of *T'* by counting the number of the entries in the scope-list of *P'*. If the percentage of such support in the total number of trees in *TDB* is more than or equal to a user specified minimum support value *minsup*, *T'* is a frequent subtree.

IV. Pattern Mining of RNA Secondary Structure

While applying the method to pattern mining of RNA secondary structure, we should turn the RNA molecule into tree-like patterns. Based on an intensive study about RNA secondary structure, we design an approach to establish the tree model for it. The elements include bulge loop, hairpin loop, internal loop, 3'-end and 5'-end of double helix and RNA-adapter are all regarded as nodes. RNA stems of complementary base pairs (more than 1) are regarded as edges. From 5' to 3' of RNA, we can orderly give the node labels with the numbering scheme

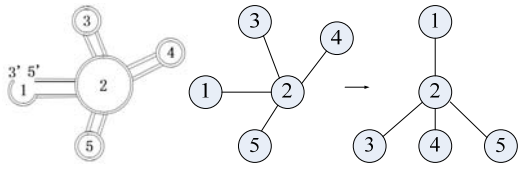


Figure 4-1 the tree model of RNA secondary structure

(1,2,…,*n*). The root is the node with label 1 and the children of the node are arranged in order of their labels. Because the 5' and 3' of RNA molecule are both fixed. Through this model establishing method, we can get a unique model for each RNA secondary structure. A labeled and ordered tree model of RNA secondary structure is given as figure 4-1:

V. Experimental Results and Analysis

A. Performance comparison on synthesis data

In this section, we conduct a set of experiments to compare the performance of the IRTM algorithm with other algorithms. All the experiments were conducted on a 3.0GHzPentium with 512MB memory. All codes were complied using Microsoft Visual C++ 6.0.

We compare the efficiency of the IRTM algorithm with PatternMatcher algorithm[8] (PM in short) and TreeMiner algorithm[8] (TM in short). The running time shown as follows includes the pretreatment time, that is to say, the time for data format conversion in TreeMiner algorithm and IRTM algorithm.

We developed a program to generate the synthetic test data. Firstly a master tree with *p* nodes and *q* node labels is constructed. Next according to parameter *t* which denotes the number of subtrees we need, the testing dataset of tree can be produced. The default setting for parameters of this procedure is as follows: p=1000, q=100, that means the nodes of tree are in the range of [0,1000], and the corresponding labels are in the range of [0,100].

In our experiments, we set some parameters as follows: the data size *T* varies from 100 to 10^6, while *minsup* keeps on 0.5. Under these conditions, the running time of PM, TM and IRTM algorithm is compared as shown in figure 5-1.

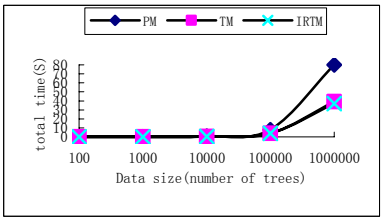


Figure 5-1 Data size vs. time

Figure 5-2 shows that with the fixed data size $T=10^4$, the running time of PM algorithm, TM algorithm and the IRTM algorithm will vary with *min_sup* increasing from 0.1 to 1. While *T* increases to 10^6 and the minimum support increases from 0.1 to 1, the running time of three algorithms are shown in Figure 5-3.

It also can be observed from these figures that algorithm IRTM is rather stable and always faster than TM and PM algorithms when varying the data quantity and the minimum support.

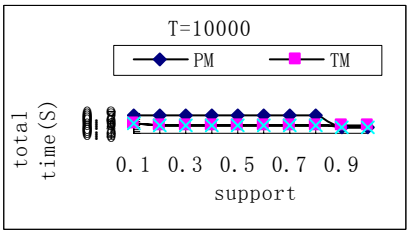


Figure 5-2 support vs.time(T=10000)

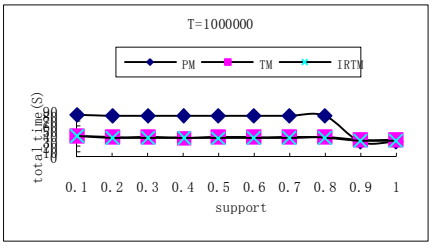


Figure 5-3 support vs. time(T=1000000)

B. Results on mining Common topological patterns of RNA molecule

We test our algorithm IRTM on 30 prokaryotic RNA structures selected from RNaseP database[11] We first establish their tree models, there are 30 trees in the database and the minimal tree size is 3 while the maximal is 13.

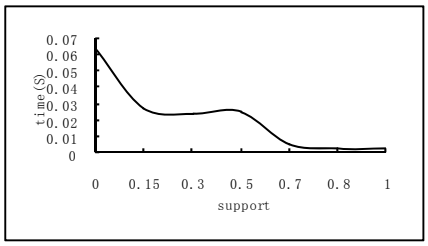


Figure 5-4 Compuation time changes with respect to the minimum support

Figure 5-4 shows the running times of the algorithm IRTM with different minimum supports. From the figure, we can see that the decreasing of minimum support will raise the number of frequent patterns satisfying with the minimum support, and the running time will grow up accordingly.

The frequent patterns numbers mined by the algorithm IRTM with the different supports are shown in figure 5-5. It is obvious from the figure that the number of frequent patterns will increase with respect to the decreasing of support. Since the frequent patterns mined are regarded as meaningless when the minimum support is set close to 0, we only observe the performance of the algorithm when the minimum support varies from 0.3 to 0.5. In this range of minimum support, the increment of patterns detected by IRTM is quite stable. This indicates the patterns mined with respect to these supports are more representative in our selected RNA secondary structures.

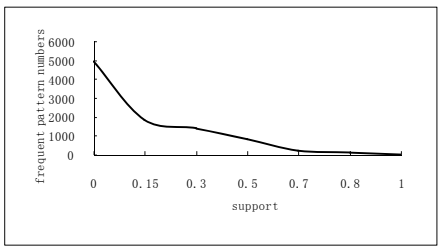


Figure 5-5 pattern numbers vs.support

Figure 5-6 shows a frequent subtree pattern mined with the support more than 9 in these 30 trees. Compared with other direct tree pattern mining algorithms, our algorithm can extract more “hidden” information and the mining results are more valuable for referential use.

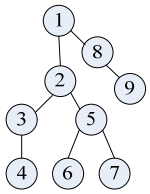


Figure 5-6 frequent tree pattern mined

VI. CONCLUSION

In this paper, we propose an efficient algorithm named IRTM for embedded frequent subtree mining of biological data, which shows significant performance improvement over previous work. We also advanced a string encoding method for representing the trees, and a scope-list to extend all substrings for frequency test. The IRTM algorithm adopts vertically mining approach and by the use of string encoding method for trees. During the process of candidate substring expansion, scope-list is used to to count their supports. After getting each substring, the IRTM algorithm uses pruning technique to greatly accelerate implementation speed.

Our analysis on the experimental results shows that in comparison with some traditional embedded subtree mining algorithms, our algorithm IRTM is more efficient, fast and stable. The new strategies we advanced here can further improve the performance of other similar algorithms.

REFERENCES

[1] Lipman D.J. , Pearson W.R. Rapid and sensitive protein similarity searches. Science, 1985: 1435-1441.

[2] Lipman D.J. , Pearson W.R. Improved tools for biological sequence comparison. Proc. Narl. Acad. Sci. 1988: 2444-2448.

[3] B.Shapiro and K.Zhang. Comparing multiple rna secondary structures using tree comparison. Computer Applications in Biosciences, 6(4):309-318, 1990.

[4] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]. In Proceedings of SIGMOD 1993, Washington, America 1993.

[5] Wang K, Liu H. Schema discovery for semistructured data[C]. In Proceedings of KDD 1997, Newport Beach, Canada.1997.

[6] Asia T, etal. Efficient substructure discovery from large semistructured dat [C]. In: Proceedings of SIAM 2002, Arlington, VA. America.2002.

[7] Chi Y, Yang Y, Muntx R R. Index and mining free trees[C]. In : Proceedings of ICDM, 2003, Melbourne, Florida, America, 2003.

[8] Zaki M J. Efficiently mining frequent trees in a forest [C]. In: Proceedings of KDD 2002, Edmonton, Alberta, Canada, 2002.

[9] Wang Chen, Hong Ming-sheng, Pei Jian, etal. Efficient pattern growth methods for frequent tree pattern mining [C]. In: Proceedings of PAKDD 2004, Sydney, Australia, 2004.

[10] Alexandre Termier, Yoshinori Tamada, Kazuyuki Numata, Seiya Imoto, Takashi Washio and Tomoyuki Higuchi. Improvement of a Gene Network Discovery Method via Frequent Subtree Mining. In Proceedings of the 17th Conference on Genome Informatics (GIW'2006), December 18-20, 2006, Yokohama, Japan.

[11] http://www.mbio.ncsu.edu/RnaseP/

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 035-038

Active Worm Propagation Modeling in Unstructured P2P Networks

Xiaosong Zhang[1], Ting Chen[1], Jiong Zheng[1] , and Hua Li[2]
[1] School of Computer Science & Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, 611731, China
Email: chenting19870201@163.com
[2] Unit 78155 of PLA

***Abstract*—Nowadays, the security of P2P networks is alarming ascribing to worms which propagate by exploiting common vulnerabilities in P2P software. Taking account of the topology of P2P networks and the behavior of worms, this paper models the propagation of active worms in unstructured P2P networks. Simulations indicate that propagation of worms in P2P networks is much faster than that in un-P2P networks. This paper highlights the analysis of the impact of worm propagation brought by attack and defense strategy changes from the angle of network topology. Finally we put forward a suggestion of worm defense in P2P networks.**

***Index Terms*—active P2P worm, modeling, network topology, attack and defense strategies**

I. INTRODUCTION

P2P networks have become latent vehicle of active worms. Recently, vulnerabilities of P2P software were advertised [1] and tended to be utilized by active worms. Ascribing to the P2P topology, propagation of worms in P2P networks is more efficient than that in un-P2P networks [2,3]. Moreover, P2P worms pose a more severe threat than traditional worms because the P2P topology upgrades the difficulty of worm detection and defense [1,3,4].

It is universally acknowledged that P2P worms could be classified into three categories: active P2P worms, reactive P2P worms and passive P2P worms. Passive P2P worms copy themselves into the share folder of the P2P client and allure other users to download these copies then complete propagation by running them in the peers' terminals [12]. Apparently, passive P2P worms cannot infect others without users' intervention. On the contrary, reactive and active P2P worms automated propagate through common vulnerabilities of P2P clients [13]. Reactive P2P worms only infect peers which are requesting files at that time while the active P2P worms aim at infecting all vulnerable nodes as quickly as possible leveraging the cached neighbors' information [13,14]. In this paper, we only discuss issues related to active P2P worms.

A P2P network is either structured or unstructured relying on its topology. All hosts maintain the same number of connections in structured P2P networks but unstructured P2P networks represent a degree distribution of power-law [4]. Since unstructured P2P networks are dominant in real-life deployment, we only discuss issues related to unstructured P2P networks in this paper.

This paper models the propagation of active worms in unstructured P2P networks. Simulations indicate that propagation of worms in P2P networks is much faster than that in un-P2P networks. The emphasis of this paper is the analysis of the impact of worm propagation brought by attack and defense strategy changes from the aspect of network topology. Finally, we give an advice of worm defense. There are three main contributions of this paper:

1) Our model is an extension of the two factor model in unstructured P2P networks.
2) Our model takes network topology and worm behaviors into account.
3) This paper reveals the impact of worm propagation brought by attack and defense strategy changes and explains from the aspect of network topology.

The rest of the paper is organized as follows. Section 2 briefly reviews the related works. Section 3 describes our model. Section 4 presents the simulations and analysis. Finally we conclude in Section 5.

II. RELATED WORKS

SEM model is the foundation of worm propagation modeling which is derived from epidemic model [5]. In SEM model, hosts can only transfer from susceptible to infectious. SIR model makes an improvement of SEM model by taking the removed state and the transition from infectious to removed into account [6]. SIS model is another evolution of SEM model assuming that infectious hosts can get back to susceptible state with certain probability [7]. Two factor model [8] may be the most comprehensive model in un-P2P networks and actually it's an extension of SIR model. Two factor model affirms that both the routers' congestion and the transition from susceptible to removed can affect worm propagation.

Models mentioned above in this section are proposed in un-P2P networks but they can be used for reference in P2P networks. Wei Yu *et al.* [4] and Zhang Yejiang *et al.* [10] transplant SEM model to P2P networks. Chaosheng Feng *et al.* [9] propose a model which is an improvement of SIS model.

III. MODELING ACTIVE P2P WORM

A. Worm Behaviors

In P2P networks, active worms adopt an optimized strategy to speed up propagation. When worms intrude into a vulnerable P2P host, neighbors of the victim are

regarded as new targets. In order to obtain an analytical model, we simplify worm behaviors as follows. Firstly, all hosts are linked to the P2P network. Secondly, once a host is infected, all its neighbors will be attacked immediately but never repeated. The last is that worm propagation from an infectious host to its neighbors only needs a unit time.

B. Model Descriptions

Since our model is an extension of two factor model in P2P networks, there are three states of hosts: susceptible, infectious and removed. Transition of these three states is represented in figure 1.

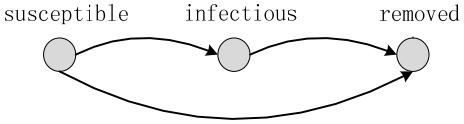


Figure 1. States transition

Our model simplifies the dynamic P2P networks to static like anther models because worms spread so fast that network topology changes are negligible. We list notations and related descriptions in table 1, which will affect the active worm propagation modeling.

TABLE I.
NOTATIONS AND DESCRIPTIONS

Notations	Descriptions
N	Total number of hosts in the P2P network, $N = S(t) + I(t) + R(t) + Q(t)$.
$S(t)$	Number of susceptible hosts at time t .
$I(t)$	Number of infectious hosts at time t .
$R(t)$	Number of removed hosts from the infectious population at time t .
$Q(t)$	Number of removed hosts from the susceptible population at time t .
$J(t)$	Number of infected hosts at time t , $J(t) = I(t) + R(t)$
γ	Average rate of removal of infectious hosts.
μ	Average rate of removal of susceptible hosts.
$V(t)$	Set of infectious hosts at time t .
d_i	Degree of host i .
σ	Average probability of worm propagation

C. Model Equations

Lemma 1 The number of infectious hosts from the susceptible population at time *t* is:

$$\sigma S(t)\left[1-(1-\frac{1}{N})^{\sum_{i\in V(t)} d_i}\right].$$

Proof: There are $\sum_{i\in V(t)} d_i$ attacks at time *t*, and given a host, the probability of been attacked by a given attack is $\frac{1}{N}$.

Then given a host, the probability of been attacked by at least one attack is $\left[1-(1-\frac{1}{N})^{\sum_{i\in V(t)} d_i}\right]$. For some reasons such as anti-virus software, firewall etc, several attacks fail to propagate worms then the probability of one given host been compromised is $\sigma\left[1-(1-\frac{1}{N})^{\sum_{i\in V(t)} d_i}\right]$.

Given a host, the probability of been a susceptible host is $\frac{S(t)}{N}$, so the result is:

$$N\times\frac{S(t)}{N}\times\sigma\left[1-(1-\frac{1}{N})^{\sum_{i\in V(t)} d_i}\right]=\sigma S(t)\left[1-(1-\frac{1}{N})^{\sum_{i\in V(t)} d_i}\right].$$

■

Like two factor model, $R(t)$ is proportion to $I(t)$ and $Q(t)$ is proportion to $S(t)\times J(t)$. $I(t)$ is increased by worm propagation but decreased by removing, so $I(t)=0$ in the end. $S(t)$ is decreased by both worm propagation and removing so finally $S(t)=0$ and $R(t)+Q(t)=N$.

Theorem 1 In a time unit, the increasing of susceptible hosts, infectious hosts and removed hosts is in equation 1. Ascribing to the paper's length limit, we do not present the proof and it can be referred to two factor model. Although two factor model in [8] is continuous and our model is discrete, proof is similar.

$$\begin{cases} I(t+1)-I(t)= \\ \sigma S(t)\left[1-(1-\frac{1}{N})^{\sum_{i\in V(t)} d_i}\right]-[R(t+1)-R(t)] \\ S(t+1)-S(t)= \\ -\sigma S(t)\left[1-(1-\frac{1}{N})^{\sum_{i\in V(t)} d_i}\right]-[Q(t+1)-Q(t)] \\ R(t+1)-R(t)=\gamma I(t) \\ Q(t+1)-Q(t)=\mu S(t)J(t) \end{cases} \quad (1)$$

IV. SIMULATIONS AND ANALYSIS

A. P2P Network Simulation

Before the simulation of worm propagation, we need to build an experimental P2P network because the commercial P2P networks are difficult to deal with. Researches show that P2P networks represent a degree distribution of power-law. BA model [11] is the first accomplishment to build a power-law network and the process is efficient. So in this paper, we build a P2P network according to the BA model.

B. Simulation Results

In this paper, we use the tuple: < N , $I(0)$, $R(0)$, $Q(0)$, μ , γ , σ > to represent the system configuration parameters. Generally, in the first stage of worm propagation, no host will be removed, so $R(0)=Q(0)=0$. At first, we set the parameters according to the two factor model as follows: <10000, 1,

0, 0, $\frac{0.8}{10000}$, 0.05, 1> and the results is demonstrated in figure 2.

We can see that $I(t)$, $Q(t)$ and $J(t)$ raise rapidly and reach the peak at the same time. After the peak, $I(t)$ falls slowly but $Q(t)$ and $J(t)$ keep stable because at the moment $S(t) = 0$.

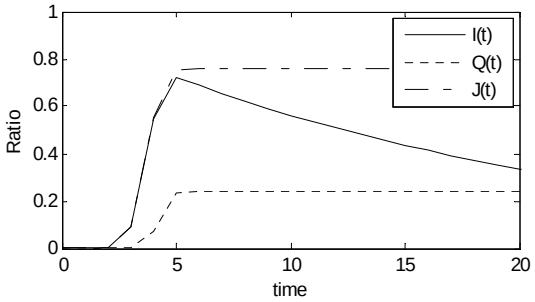


Figure 2. Worm propagation Trend of Our Model

C. Comparison with Other Models

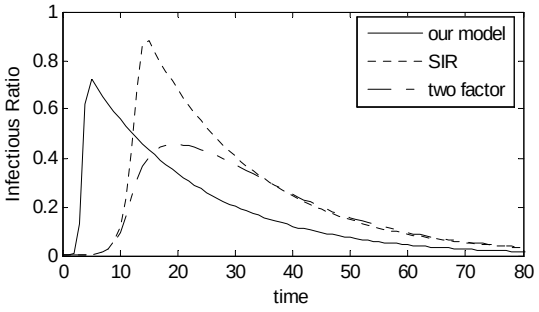


Figure 3. Comparison of our model with SIR model and two factor model

Figure 3 gives a comparison of our model with SIR model and two factor model. We find that worm propagation in P2P networks is much faster than that in un-P2P networks. Figure 3 validates that worms in P2P network pose a more severe threat than those in un-P2P network because the maximum of $I(t)$ of our model is much higher than that of two factor model. It's rational that the peak of SIR model is higher than that of our model because SIR model does not take the transition of susceptible hosts to removed into account.

D. The Sensitivity of Attack/Defense Strategy

Since this paper focus on the analysis of the impact of worm propagation brought by attack and defense strategy changes from the angle of network topology, here we do not give the results of worm propagation with other parameters changes.

Figure 4 examines the impact brought by attack strategy changes. Random strategy means that the initial victims are selected randomly. While target strategy denotes that worms infect the most connected hosts in the beginning. We find that the peaks of the two curves are more or less the same and the tails of them are overlapped. The reason is that $R(t)$, $I(t)$ and $Q(t)$ is proportioned. Moreover before the two curves reach the peak, target strategy keeps ahead of random strategy about two unit times. The reason from the network topology is shown in table 2: the average distance from a random host to all other hosts is approximately two larger than that from a host with the largest degree.

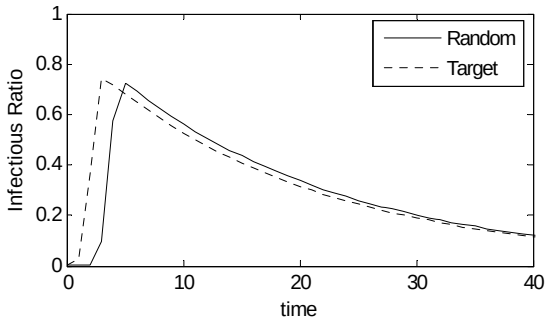


Figure 4. The sensitivity of attack strategy

TABLE II.
DISTANCES OF DIFFERENT STRATEGIES($N = 10000$)

Random	Target
4.65	2.59

Figure 5 reveals the impact brought by defense strategy changes. Random strategy means that the initial removed hosts are selected randomly. While target strategy denotes that we immunize the most connected hosts in the beginning. It's obviously that worm threats can be alleviated by preferential immunizing frequently connected hosts before worm propagation. In this experiment, the number of initial removed hosts Q_0 is a variable and we use two different defense strategies: one is to immunize random nodes and the other is to immunize the most connected nodes before worm propagation. From figure 5, we find that random immunization has little effect to improve the global security: worms can infect almost the other nodes. On a contrary, worm propagation can be constrained by immunization a few number of the most connected nodes: about 20% nodes in figure 5.

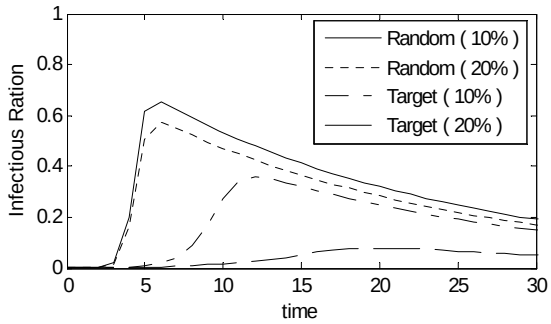


Figure 5. The sensitivity of defense strategy

We present the reason that target defense is better than random defense from the aspect of the network topology. There are many edges connect to the most connected nodes, if these nodes are removed, all edges connected to

these nodes are also removed, the connectivity of P2P networks decreases a lot. However, the global connectivity changes little when remove a few of the less frequently connected nodes because they possess a small quantity of edges. Figure 6 depicts the change of network diameters under random removing and target removing of the most connected nodes. The term diameter is defined as the average length of the shortest paths between any two nodes in the network. Figure 6 shows that P2P networks are stable by random removing but vulnerable by target removing. With the increasing of the diameter of P2P network, worm propagation remarkablely slows down. When the number of removed hosts beyond a threshold, the P2P network breaks into several isolate subnets then worms are constrained in their subnets. So our suggestion of worm defense is that protecting critical nodes of the P2P network from compromising.

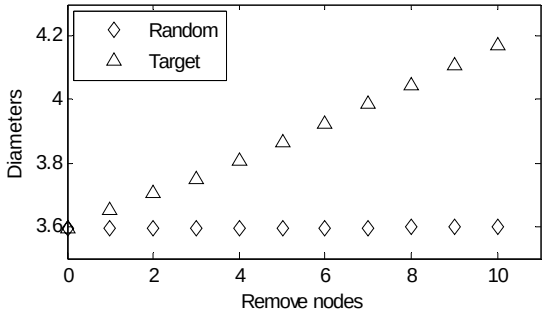


Figure 6. Diameter trend of P2P network ($N = 2000$)

Your goal is to simulate the usual appearance of papers in a Conference Proceedings of the Academy Publisher. We are requesting that you follow these guidelines as closely as possible.

V. CONCLUSION

This paper models the propagation of active worms in unstructured P2P networks. Our model takes network topology and worm behaviors into account. Moreover, it is an extension of the two factor model in unstructured P2P networks. This paper focus on the impact of worm propagation brought by attack and defense strategy changes and explains from the aspect of network topology. In the end, we give an advice of worm defense is that protecting critical nodes of the P2P network from compromising.

ACKNOWLEDGMENT

This work was supported in part by a grant from Science and Technology Commission of Shanghai Municipality 09511501600.

REFERENCES

[1] random nut. The PACKET 0' DEATH FastTrack network vulnerability. NETSYS.COM Full Disclosure Mailing List Archives, May 2003. http://www.netsys.com/fulldisclosure/2003/05/msg00351.html.

[2] L. Zhou, L. Zhang, F. McSherry, N. Immorlica, M. Costa and S. Chien, "A First Look at Peer-to-Peer Worms: Threats and Defenses," In Proceedings of Peer-to-Peer Systems IV, 4th International Workshop (IPTPS), pages 24-35, February 2005.

[3] W. Yu, S. Chellappan, X. Wang, and D. Xuan, "On Defending Peer-to-Peer System-based Active Worm Attacks", Proceedings of 2005 IEEE Global Telecommunications Conference, IEEE Press, Piscataway,2006, pp. 1757-1761.

[4] Yu, Wei; Chellappan, Sriram; Wang, Xun; Xuan, Dong. Peer-to-peer system-based active worm attacks: Modeling, analysis and defense. Computer Communications, v 31, n 17, p 4005-4017, November 20, 2008.

[5] Zou CC, Gong W, Towsley D. On the performance of Internet worm scanning strategies. Technical Report, TR-03-CSE-07, Electrical and Computer Engineering Department, University of Massachusetts, 2003.

[6] Frauenthal JC. Mathematical Modeling in Epidemiology. New York: Springer-Verlag, 1980.

[7] Wang Y, Wang CX. Modeling the effects of timing parameters on virus propagation. In: Staniford S, ed. Proc. of the ACM CCS Workshop on Rapid Malcode (WORM 2003). Washington, 2003.

[8] Zou CC, Gong W, Towsley D. Code Red worm propagation modeling and analysis. In: Proc. of the 9th ACM Symp. on Computer and Communication Security. Washington, 2002. 138~147.

[9] Feng Chaosheng, Qin Zhiguang, Cuthbet Laurence, Tokarchuk Laurissa. Propagation model of active worms in P2P networks. In Proceedings of the 9th International Conference for Young Computer Scientists, ICYCS 2008, p 1908-1912.

[10] Zhang Yejiang, Li Zhitang, Hu Zhengbing, Huang Qingfeng, Lu Chuiwei. Evolutionary proactive P2P worm: Propagation modeling and simulation. In Proceedings - 2nd International Conference on Genetic and Evolutionary Computing, WGEC 2008, p 261-264.

[11] Albert R, and Barabási A L, Statistical Mechanics of Complex Networks. Rev. Mod. Phys, 2002, pp. 47-97.

[12] R. Thommes and M. Coates, "Epidemiological Modeling of Peer-to-Peer Viruses and Pollution," In Proceedings of IEEE INFOCOM, April 2006.

[13] G. Chen and R.S. Gray, "Simulating non-Scanning worms on peer-to-peer networks," Proc. ACM Conf. Scalable Information Systems (INFOSCALE 06), ACM Press, May. 2006, pp. 29-41,doi:10.1145/1146847.1146876.

[14] Zhitang Li, Yejiang Zhang, Zhengbing Hu, Huaiqing Lin, and Chuiwei Lu. Network-Based Detection Method against Proactive P2P Worms Leveraging Application-Level Knowledge. In Proceedings of 2009 First International Workshop on Education Technology and Computer Science. 2009, pp.575-580.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 039-044

Method of the Object-oriented Program Exact Testing

Xiaolan Wang [1], Yanshuai Zhang [2], and Hong He [3]
[1] Department of Information and Engineering of Shandong University at Weihai, Weihai, China
Email: wangxiaolansdu@163.com
[2] Department of Information and Engineering of Shandong University at Weihai, Weihai, China
Email: zys12345678@126.com
[3] Department of Information and Engineering of Shandong University at Weihai, Weihai, China
Email: hehong@sdu.edu.cn

***Abstract*—Object-oriented programming Exact Testing is an important research direction on testing. But there haven't been any effective methods of error tracking and positioning in object-oriented programming exact testing. In this paper, a method, based on symbolic execution and constraint solving, is proposed to build the dependency graph of the error statement. Compared with other studies, this method is more accurate. Experiments show that this method can be used to track and position error in testing procedures for small and medium-size process. It has a wide application in many areas, such as program testing, debugging and code optimization, etc.**

***Index Terms*—Object-Oriented program, Error Tracking, Symbolic Execution, Exact Testing**

I. INTRODUCTION

Correctness is one of the most important attributes of program. Ensuring the correctness of program and finding the implicit mistakes as far as possible are always an important problem focused in world of computer science.

With the development of programming languages, an increasing number of software development methods have been transformed form structured-oriented into the object-oriented. On that account, object-oriented program testing is increasingly significant.

In this century, there are plenty of domestic and foreign scholars who throw themselves into object-oriented program testing. For instance, Ugo Buy uses data flow analysis and symbolic execution to produce a series of methods to call CUT[2]. Paolo Tonella uses genetic algorithm to generate test cases automatically，then these cases would be used in unit testing for classes[3]. Minh Ngoc Ngo and Hee Beng Kuan Tan summarized the characteristics of infeasible path in common programs by observing[4]. They proposed heuristics-based infeasible path detection for dynamic test data generation. However, they were all directed at a single class testing of object-oriented program. Furthermore, explains for the practical implementation of test cases and how to deal with the errors in testing process were not given.

In this paper, we combine existing achievements in object-oriented testing with symbolic execution and constraint solving proposed by Zhang Jian. Problems that jTGEN^{I} cannot deal with by its method of determining infeasible path could be resolved. This enables jTGEN^{I} to be more accurate. As a result, jTGEN^{I} would be more effective and reliable in object-oriented program testing. In addition, this paper proposed methods of error tracking and positioning in object-oriented program testing.

This paper is organized as follows: Section II summarizes two methods of infeasible path determining which would be used in the following sections. Section III introduces the infeasible path determining method which is based on symbolic execution and constraint solving, heuristics-based infeasible path detection used by Ngo and Tan, and test data generation tool jTGEN^{I} which is based on symbolic execution and constraint solving. Section IV describes error tracing methods and related algorithms and positioning mode. Section V proves the validity and correctness of the method proposed in this paper by examples. Section VI is conclusion.

II. INFEASIBLE PATH DETERMINE METHODS

A. Accurate determine method based on symbolic execution and constraint solving

The method based on symbolic execution and constraint solving is accurate.

Symbolic execution [5] is using symbol as a value to bestow on the variable and imitate the execution of path. If the initial value of variable 'i' is symbol 'a', after 'j=i+1', then the value of 'j' would be 'a+1'.

Through symbol execution, a set of constraints on the initial value of the variable could be obtained, that is the path condition.

B. Heuristics-based infeasible path detections

Ngo and Tan got some common features of infeasible paths. They thought that the main reason of programs containing infeasible paths is that there are interrelated conditional statements. Then they advanced heuristics-based concepts of determining conditional statements. They defined the attributes of infeasible paths and proved them. What's more, they applied the technique into a new test case generation tool jTGEN^{I}.

C. Comparison of two methods

Between the above two methods that determine the feasibility of paths, method based on symbolic execution and constraint solving is more accurate than heuristics-based infeasible path detection. And the former one is much easier in execution. However, it only applies to small and medium-scale program as a result of computational complexity becomes bigger with the increasing number of variables.

III. JTGEN^{SI} BASED ON SYMBOLIC EXECUTION AND CONSTRAINT SOLVING

A. Introduction of jTGEN^{I}

jTGEN^{I} is semi-automatic path-oriented test cases generation tool. Letter 'j' stands for java, and the superscript 'I' stands for infeasible path detection. Infeasible path detection could be inner-process or inter-process. jTGEN^{I} implements the detection inner-process. Figure 1 shows the structure diagram of jTGEN^{I}.

As a path-oriented test case generation tool, the fifth part of jTGEN^{I} is an infeasible path detector. And its determining method of infeasible path is inaccurate.

B. jTGEN^{SI} based on symbolic execution and constraint solving

Infeasible path detection accuracy of jTGEN^{I} is over 90%. But it still cannot reach "precision". After symbolic execution and constraint solving method takes the place of heuristics-based infeasible path detection, the infeasible path detection part of jTGEN^{I} would be much more accurate. The method of symbolic execution and constraint solving is running a program after a symbolic substitute for the variable. Then we would obtain a set of constraints on the symbols. The problem of constraint satisfaction could be resolved by constraint solving algorithm. At length, whether the path is feasible or not could be determined. In order to distinguish it from jTGEN^{I}, we call it jTGEN^{SI} that is based on symbolic execution and constraint solving. Following we proof that, with symbolic execution and constraint solving method, we can make the testing object-oriented programming become more precise.

C. Examples of the accuracy of method based on symbolic execution and constraint solving

Example 3.1

```
if( o instanceof Circle){
((Circle) o).getS(r);
}
If(o instanceof Rect){
((Rect) o).getArea(w,h);
}
```

This piece of code is a part of the program that selects different squaring methods for different graphs. Following we use the method in [6] to analysis it.

```
int r,w,h;
boolean b1= (o instanceof  Circle);
boolean b2= (o instanceof Rect);
1 @ b1==true;
2   s = o.getS(r);
3 @ b2==true ;
4   s = o.getArea(w,h);
```

'@' make a distinction between logical expressions and assignment statements.1~4 state clearly about the implementation order of the statements in a path.

1.Constraints on the symbols

b1= (o instanceof Circle);

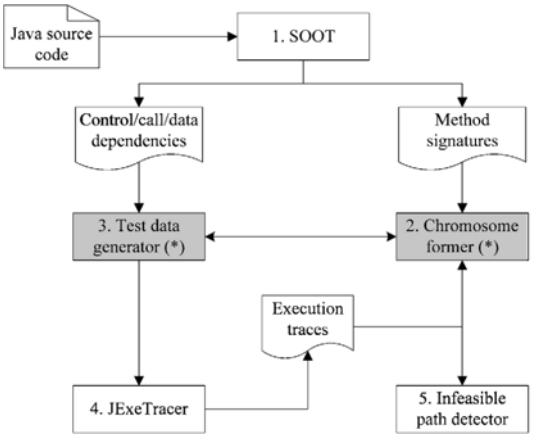


Figure1. The structure diagram of jTGEN^{I}

b2= (o instanceof Rect);

2. Program path

Path1： 1-2-3-4

@ b1==true; s = o.getS(r);

@ b2==true ; s = o.getArea(w,h);

Path2 ： 1-2-3

@ b1==true; s = o.getS(r); @ b2==false ;

Path3 ： 1-3-4

@ b1==false; @ b2==true ; s = o.getArea(w,h);

Path4 ： 1-3

@ b1==false; @ b2==false ;

3. Analysis results of each path

Path1: o instanceof Circle is true and o instanceof Rect is true.

By Definition 1, when 'o instanceof Circle' is true, 'o instanceof Rect' cannot be true. The constraints of path1 wouldn't be satisfied. Thus path 1 is infeasible.

Path2: o instanceof Circleis true and o instanceof Rect is false.

By Definition 1, the constraints of path 2 could be satisfied. Thus path 2 is feasible.

Path3: o instanceof Circle is false and o instanceof Rect is true.

By Definition 1, the constraints of path 3 could be satisfied. Thus path 3 is feasible.

Path4: o instanceof Circle is false and o instanceof Rect is false.

By Definition 1, the constraints of path 4 could be satisfied. Thus path 4 is feasible.

This piece of code only has one infeasible path, three feasible paths. Therefore, with symbolic execution and constraint solving in its infeasible path detection part, jTGEN^{SI} will generate test cases more effective and accurate.

IV. TRACKING AND POSITIONING ERRORS

A. Summary

During the test, when the actual results and expected results are not same, testers often have to find out the errors. Tracking the errors accurately and quickly will reduce the workload of testers and improve the testing efficiency when the program is complicated and codes are numerous. This chapter combines symbolic execution and constraint solving with dependency determination algorithm to generate dependence graph of the error statement. Then according to tracking segmentation integrated method, the position of error will be fixed.

B. Generation of the dependence graph

1. Method of dependency determination

Definition 1[8] S_1 and S_2 are two statements in program P , V is a variable if $V \in DEP(S_1)$ and $V \in REP(S_2)$. And S_2 uses the value of V that is calculated in S_1 when P executes from one path. We call that S_2 is (directly) dependent on S_1 concerning variable data stream.

Theorem 1[8] S_1 and S_2 are two statements in program P , V is a variable. The relationship $DEP_v(S_x, S_y, V)$ could be established only when:

$path(S_1, S_2, ..., S_n)$ is exited and $s_1 = s_x, s_n = s_y$;

$DEP_{pv}(S_i, S_n, V, path(S_1, S_2, ..., S_n))$,

$(i = 2,3,...,n-1)$ is unsubstantiated. That means $V \notin DEP(S_i)$.

2. Elements of dependency graph

(1) The node of dependent statement list

```
Struct D_Node{
  Struct Statement * s //Point to the dependent statement
  Struct D_Node * next;// Point to the next dependent node
};
```

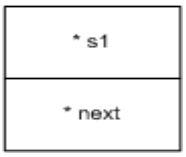


Figure 2. Structure of dependent node

(2) Statement

```
Struct statement{
  Struct D_Graph * p; //Point to the path the statement belongs
  Struct D_Node * d; //Point to the head of the list which saves all the dependent nodes of current statement
  Struct statement * s'; // Points to the next statement of the path
  String class_Name; // Name of the file
  int line;//The number of row
  Struct p * p'; //Point to the feasible path set of the function that statement calls from the internal or outer class
};
```

(3) Path

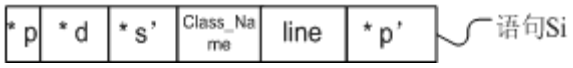


Figure 3. Structure of statement

```
Struct p{
Struct Statement * p;//Point to statement
};
```

3. Generation Algorithm of dependency graph of the error statement

Generation Algorithm of dependency graph of the error statement is represented as follows:

Input the error statements discovered in sequence $s_1, s_2, ..., s_m$

For i =1 to m

Step1: Call PAT in the function $f_i( )$ where the error statement S_i is found to get the path set $P[n]$.

// $P[i]$ is a list. Its node is statements the path contains. Each statement point to the next on with the pointer s'

Step2：For each statement in $P[i]$, if($p' \neq null$), then p point to the path set in this function;

// p' is a pointer. It points to the feasible path set of the function which is called.

Step3: Create dependency node for the statements $S_q, S_K, \cdots, S_n$ which S_i depend on in $P[i]$.

//The directly dependent graph of the error statement in $P[i]$ could be obtained. We can directly get the filename, the number of line of each statement and the dependent statements of the error statement.

Step4: if (S_j calls other function $f_j( )$), call PAT in $f_j( )$ to get its path set $p'[n]$, repeat Step1、Step2、Step3;

Output the dependency graph $G_1, G_2, \cdots G_m$ that the error statements sequence within the whole program.

The time complexity of this algorithm is $O(m*n)$, m is the quantity of error statements, n is the quantity of statements that the path contains. The complexity would never be over $O(n^2)$. Fig.4 shows dependency graph of S_n in single process.

C. Positioning error

This paper uses a method that integrated tracking and segmentation to locate error. This method combines the advantages of tracking and segmentation. Positioning error is fast and its security is good. Therefore it is more suitable for large debugging. Suppose the program includes processes P_0 , P_1 , $\cdots$ P_n , P_0 is the main process.It accomplishes inputting data, calls other

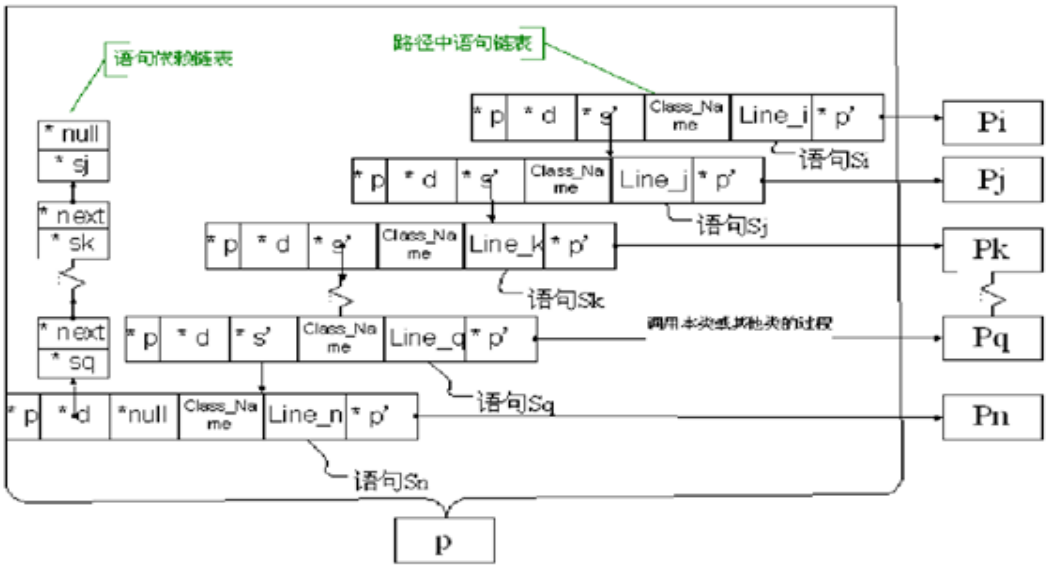


Figure 4. Dependency graph of S_n in single process.

programs and output the results of variables. If there are some statements which impact the variable 'V' in P_0, P_1, $\cdots$ P_n and there is an error in P_3. When P_0 output a wrong value of V, the error region would be large, with segmentation used. Because in every process, errors may be in statements which the variable V relates to However, if tracking segmentation integrated method is used, the error region could be reduced greatly. Therefore the error could be located quickly. The main idea of tracking segmentation integrated method is: First using forward tracking can reduce the error region. That means it would restrict the error region in the process P_3. Next use dynamic segmentation on the variable which has the wrong value. Last locate the error in the segment.

V. EXPERIMENT AND ANALYSIS

```
class Area {
    private int width;
    private int height;
    private int radius;
    public int squarenessArea(Area area) {
        int width = area.getwidth();
        int height = area.getheight();
        return width * height;
    }
    public int circleArea(Area area) {
        int radius = area.getRadius();
        return (int) ((3.1415926) * radius * radius);
    }
    public int getheight() {
        return height;
    }
    public void setheight(int height) {
        if (height > 0)
            this.height = height;
        else {
            System.out.println("输入的数值应大于零");
            System.exit(0);
        }
    }
    public int getRadius() {
        return radius;
    }
    public void setRadius(int radius) {
        if (radius > 0)
            this.radius = radius;
        else {
            System.out.println("输入的数值应大于零");
            System.exit(0);
        }
    }
    public int getwidth() {
        return width;
    }
    public void setwidth(int width) {
        if (width > 0)
            this.width = width;
        else {
            System.out.println("输入的数值应大于零");
            System.exit(0);
        }
    }
}
```

Figure 5. AreaClass.java

```
class AreaOperation {
    private int totalArea;

    public void compareArea(int firstArea, int secondArea) {
        if (firstArea > secondArea)
            System.out.println("第一个的面积大于第二个的面积");
        else if (firstArea < secondArea)
            System.out.println("第一个的面积小于第二个的面积");
        else if (firstArea == secondArea)
            System.out.println("第一个的面积等于第二个的面积");
    }

    public void addAreas(int firstArea, int secondArea) {
        this.totalArea = firstArea + secondArea;
        System.out.println("面积和为: " + totalArea);
    }
}
```

Figure 6. AreaOperation.java

```
class MainClass {
    public static void main(String argv[]) {
        Area area = new Area();
        AreaOperation areaOperation = new AreaOperation();
        area.setwidth(600);
        area.setheight(40);
        area.setRadius(20);
        int squarenessArea = area.squarenessArea(area);
        int circleArea = area.circleArea(area);
        areaOperation.compareArea(squarenessArea, circleArea);
        areaOperation.addAreas(squarenessArea, circleArea);
    }
}
```

Figure 7. MainClass.java

A. *A java program*

1. A java program

The AreaClass.java, AreaOperation.java and MainClass.java can be seen in Fig.5, Fig.6 and Fig.7.

2. Analysis

In order to prove the effectiveness of the method described in this paper, we created an error deliberately. The twelfth line in Are.java is changed from " return (int) ((3.1415926) * radius * radius); " to "return (int) (2*(3.1415926)* radius);"

Following we analyze the program according to algorithm to generate dependence graph of the error statement.

Where the error occurred is the eleventh line in MainClass.java. There is only on path that passes the error statement in main(). The path is: 3-4-5-6-7-8-9-10-11;

The eleventh error statement depends on the eighth and ninth statement. Then we can create the statement dependency list and statement list in path. Figure 8 shows

the statement dependency list of MainClass: 11 and its path diagram.

We know that the eleventh statement calls addArea() in class AreaOperation from Fig.8.

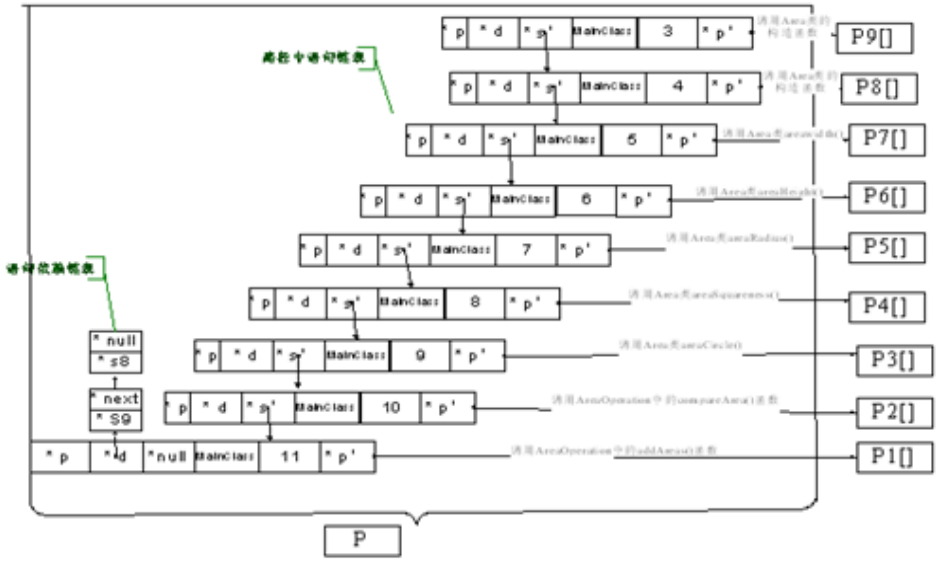


Figure 8.The statement dependency list of MainClass: 11 and its path diagram.

Then use symbolic execution and constraint solving method in this function. We can obtain the feasible path of this function: AreaOperation.java:14-15.

Now we analyze the dependent relationship of MainClass.java:8 and MainClass.java:9 according to algorithm.

Dependent relationship analysis of MainClass: 9 is showed by Fig.9.

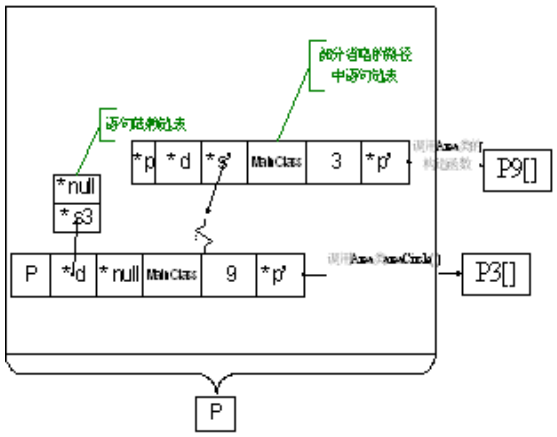


Figure 9. Dependent relationship analysis graph of MainClass: 9

We know that the ninth statement calls areaCircle () in class Area from Fig.9. And there is only one path in this function: Area.java:11-12. MainClass: 9 is dependent on MainClass: 3. MainClass: 3 calls construct function of Area. In this program, the construct function is empty.

Dependent relationship analysis of MainClass:8 is showed by Fig.10.

We know that the eighth statement calls addSquareness () in class Area. There is only one path in this function: Area.java:6-7-8. Area.java:6 calls getWidth() of Area. Area.java:6 calls getHeigth() of Area.

MainClass.java:8 is dependent on MainClass: 3. MainClass: 3 calls construct function of Area. In this program, the construct function is empty. Now, we obtain the whole statement dependence graph of the error

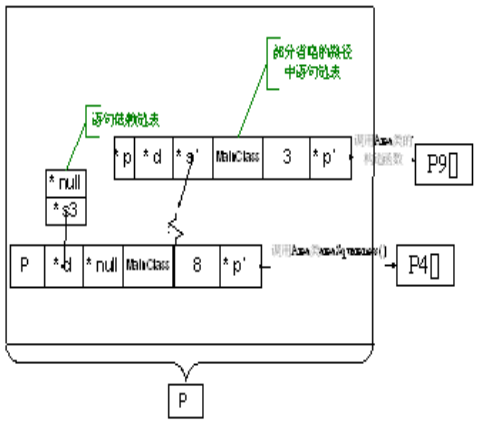


Figure 10. Dependent relationship analysis graph of MainClass:8

statement. It could be seen in Fig.11.

We combine symbolic execution with tracking and segmentation integrated method to locate error in the example. With forward tracking way, we set up a checkpoint at each call point. Compare the actual variable value obtained with the correct value at checkpoint. For instance, input width=a, height=b, radius=r. At the checkpoint of circleArea(), the actual value is 2*r*(3.1415926) which is inconsistent with the expected value r*r*(3.1415926). Then we can definite that the error is in circleArea(). The range of error is the statements int radius = area.getRadius(); return (int) (2*(3.1415926) * radius) in Area.java;

Then we analyze the function by dynamic segmentation. We can position the error rapidly because there are only two lines in it. The line where error occurs is Area.java:12. The statement is return (int) (2*(3.1415926) * radius). The right way should be return (int) ((3.1415926) * radius * radius).

In general, Area.java:12 accounts for the error of MainClass.java:11.

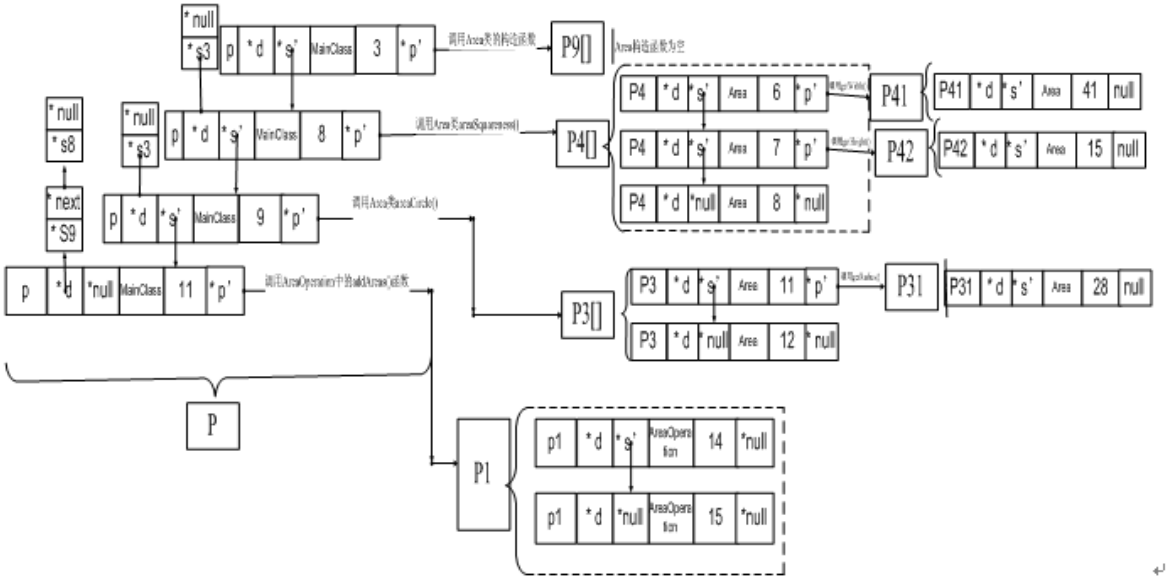


Figure 11.The statement dependency list of MainClass: 11 and its path diagram

VI. Conclusion

Object-oriented programming exact testing is one of the difficult problems of testing. This paper applies symbolic execution and constraint solving to object-oriented programming exact testing. Improve and perfect the accuracy testing tools now available. Test cases would be more reliable. It also lightens the testers' burden and raises their working efficiency. Besides, with infeasible path detection and segmentation integrated method, this paper create an accurate method to track and locate error in object-oriented programming testing. This method could be used in many systems and it can detect errors in different languages.

Although the method proposed in this paper is accurate, but there are still some limitations. For example, symbolic execution and constraint solving would be very complicated in large-scale programs. During building the dependency graph, it may cost amount of time when analyzing right statements, which cause a lot of redundancy.

How to solve the above problems is the direction of the continue study on this subject.

References

[1] Hoare C A R. The verifying compiler: A grand challenge for computing research. Journal of the ACM, 2003, 50(1):63-69.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] U Buy, A. Orso, and M. Pazze. Automated testing of classes. In Proceedings of the International Symposium on Software Testing and Analysis(ISSTA 2000) Portland, OR, USA, August 2000.

[3] P. Tonella, Evolutionary testing of classes, presenter=d at ACM SIGSOFT International Symposium on Software Testing and Analysis, ISSTA 2004, July 11-14 2004, Boston, MA, United States, 2004.

[4] Ngo M N, Tan H B K. Heuristics-based infeasible path detection for dynamic test data generation. Information & Software Technology, 2008, 50(7-8):641-655.

[5] King J C. Symbolic execution and testing. Communications of the ACM, 1976, 19(7):385-394.

[6] Zhang J, Wang X. A constraint solver and its application to path feasibility analysis. International Journal of Software Engineering and Knowledge Engineering, 2001, 11(2): 139-156.

[7] ZHANG Jian. Sharp Static Analysis of Programs. Chinese Journal of Computers, 2008，31(9) :1549.

[8] Xu Baowen. REVERSE PROGRAM FLOW DEPENDENCY ANALYSIS AND ITS APPLICATIONS. Chinese Journal of Computers, 1993, 16(5) :385.

[9] Xu Baowen. Dependence structure analysis-based approach for measuring importance of classes. JOURNAL OF SOUTHEAST UNIVERSITY(NATURAL SCIENCE EDITION),2008，38(3) :380.

[10] SHI Jun. Methods of positioning errors in debugging. Microcomputer applications. 1998, 19(5).

[11] Zhang J et al. Path-Oriented Test Data Generation Using Symbolic Execution and Constraint Solving Techniques//Proceedings of 2nd International Conference on Software Engineering and Formal Methods(SEFM 2004). Beijing, China, 2004:242-250.

[12] Zhang J. Constraint solving and symbolic execution//Proceedings of the VSTTE. Zurich, Switzerland, 2005:539-544.

[13] He Hong, Ma Shaohan. *Algorithm Analysis and Design Technology*, Beijing: Science Press, 2004.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 045-048

Aggregated Binary Relations and Rough Approximations

Liping An [1], and Lingyun Tong [2]
[1] Business School, Nankai University, Tianjin 300071, China
Email: anliping2000@sina.com
[2] School of Management, Hebei University of Technology, Tianjin 300130, China
Email: tonglingyun2008@sina.com

***Abstract*—The core concepts of classical rough sets are lower and upper approximations based on indiscernibility relations. In some cases, it is necessary to generalize indiscernibility relation by using some other binary relations. We first consider the indiscernibility relations defined from nominal attributes, similarity relations defined from quantitative attributes, and outranking relations defined from ordinal attributes. These relations defined by single attributes are aggregated into the global relations at the level of the set of attributes. The lower approximation and the upper approximation of the upward and downward unions of decision classes are defined using the global relations. Then, we investigate some of common properties of the relation based lower and upper approximation operations.**

***Index Terms*—binary relations, rough sets, lower and upper approximations, similarity, outranking**

I. Introduction

The rough set theory [1, 2] is an extension of set theory for the study of intelligent systems characterized by insufficient and incomplete information. The classical rough set theory is based on indiscernibility relations. The indiscernibility classes are the building blocks for the construction of the lower and upper approximations. However, as pointed out in [3], the rough set theory built on a partition induced by indiscernibility relations may not provide a realistic view of relationships between elements in real-world application.

Different kinds of generalizations of the classical rough set model can be obtained by replacing the indiscernibility relation with an arbitrary binary relation. Papers [4-7] have done extensive research on binary relation based rough sets. These authors started from the properties of binary relations to investigate the essential properties of the lower and upper approximation operations generated by such relations. Greco et al.[8, 9] introduced the rough approximation based on dominance relations and proposed a rough set methodology to analyze multi-criteria choice and ranking decision problems.

In this paper, we construct the indiscernibility relation for the subset of nominal attributes; the outranking relation for the subset of ordinal attributes; and the similarity relation for the subset of quantitative attributes. Then the binary relations defined are aggregated into a global relation. Set approximation can be defined based on the global relations. Some of common properties of the relation based lower and upper approximation operations are investigated.

II. Construction of Global Binary Relations

The standard rough set model can be generalized by considering any type of binary relation on attribute values instead of the indiscernibility relation.

Definition 1 [10]**.** Let U be a non-empty finite set. $U\times U$ is the product set of U and U. Any subset R of $U\times U$ is called a binary relation on U. For any $(x, y)\in U\times U$, if $(x, y)\in R$, we say x has relation R with y, and denote this relationship as xRy, i.e., $R=\{(x, y): xRy\}$.

Definition 2 [11]. Let U be a universe of discourse, $R\subseteq U\times U$ is a binary relation on U. The relation R is said to be serial if there exists $y\in U$ such that $(x, y)\in R$ for all $x\in U$; R is said to be reflexive if $(x, x)\in R$ for all $x\in U$; R is said to be symmetric if for all $x, y\in U$, $(x, y)\in R$ implies $(y, x)\in R$; R is said to be transitive if for all $x, y, z\in U$, $(x, y)\in R$ and $(y, z)\in R$ imply $(x, z)\in R$; R is said to be antisymmetric if for all $x, y\in U$, $(x, y)\in R$ and $(y, x)\in R$ implies $x=y$; R is said to be complete if $(x, y)\in R$ or $(y, x)\in R$ for all $x\neq y$.

Definition 3. The indiscernibility relation I_P on U is defined as follows:

$$I_P=\{(x, y)\in U\times U: yI_ax, \forall a\in P^{=}\}.$$

Definition 4. The indiscernibility class of an object $x\in U$ with respect to P, denoted by $I_P(x)$, is the set of objects which are indiscernible to x on P:

$$I_P(x)=\{y\in U: yI_Px\}.$$

Indiscernibility relation is an equivalence relation, which is reflexive, symmetric and transitive.

Definition 5. The outranking relation $D_P^{\succeq}$ on U is defined as follows:

$$D_P^{\succeq}=\{(x, y)\in U\times U: x\, D_b^{\succeq} y, \forall b\in P^{\succeq}\}.$$

Definition 6. The outranking class of an object $x\in U$ with respect to $P^{\succeq}$, denoted by $D_P^{\succeq}(x)$, is the set of objects which are outranking x on $P^{\succeq}$:

$$D_P^{\succeq}(x)=\{y\in U: y\, D_P^{\succeq} x\}.$$

Definition 7. The outranked relation of $P^{\succeq}$ on U, denoted by $D_P^{\preceq}$, is defined as follows:

$$D_P^{\preceq}=\{(x, y)\in U\times U: y\, D_b^{\succeq} x, \forall b\in P^{\succeq}\}.$$

Definition 8. The outranked class of an object $x \in U$ with respect to $P^{\succeq}$, denoted by $D_P^{\preceq}(x)$, is the set of objects which are outranked by x on $P^{\succeq}$:

$$D_P^{\preceq}(x)=\{y \in U: x\, D_P^{\succeq}\, y\}.$$

Outranking relation is reflexive and transitive.

Similarity relations should not be imposed to be symmetric[12, 13]. In such cases, we can consider the left neighborhood and the right neighborhood of an object in a similarity relation [10].

Definition 9. Let $R \subseteq U \times U$ be a binary relation on U. For any $x \in U$, we call the set $R^l(x)=\{y \in U: yRx\}$ the left neighborhood of x in R.

Definition 10. Let $R \subseteq U \times U$ be a binary relation on U. For any $x \in U$, we call the set $R^r(x)=\{y \in U: xRy\}$ the right neighborhood of x in R.

Definition 11. The similarity relation S_P is defined as follows:

$$S_P=\{(x, y) \in U \times U: xS_c y, \forall c \in P^{\sim}\}.$$

Definition 12. The similarity class of an object $x \in U$ with respect to $P^{\sim}$, denoted by $S_c^l(x)$, is the set of objects which are similar to x on c:

$$S_P^l(x)=\{y: ySPx\}.$$

Definition 13. The class of objects to which x is similar with respect to $P^{\sim}$, denoted by $S_c^r(x)$, is:

$$S_P^r(x)=\{y: xS_P y\}.$$

Several binary relations, such as indiscernibility, similarity and outranking relations, can be considered jointly.

Definition 14. Some global binary relations of P on U is defined as follows:

$$R_P^{l\succeq}=\{(x, y) \in U \times U: yI_P x \wedge y\, D_P^{\succeq}\, x \wedge yS_P x\},$$

$$R_P^{r\succeq}=\{(x, y) \in U \times U: yI_P x \wedge y\, D_P^{\succeq}\, x \wedge xS_P y\},$$

$$R_P^{l\preceq}=\{(x, y) \in U \times U: yI_P x \wedge y\, D_P^{\preceq}\, x \wedge yS_P x\},$$

$$R_P^{r\preceq}=\{(x, y) \in U \times U: yI_P x \wedge y\, D_P^{\preceq}\, x \wedge xS_P y\}.$$

Proposition 1. We have the following statements:

$$R_P^{l\succeq}(x)=I_P(x) \cap D_P^{\succeq}(x) \cap S_P^l(x),$$

$$R_P^{r\succeq}(x)=I_P(x) \cap D_P^{\succeq}(x) \cap S_P^r(x),$$

$$R_P^{l\preceq}(x)=I_P(x) \cap D_P^{\preceq}(x) \cap S_P^l(x),$$

$$R_P^{r\preceq}(x)=I_P(x) \cap D_P^{\preceq}(x) \cap S_P^r(x).$$

III. ROUGH APPROXIMATIONS

The sets to be approximated are called the upward union and downward union of decision classes, respectively [8]:

$\mathrm{Cl}_t^{\succeq}=\bigcup_{s \geq t} \mathrm{Cl}_s$, $\mathrm{Cl}_t^{\preceq}=\bigcup_{s \leq t} \mathrm{Cl}_s$, $t=1, 2, \ldots, n$.

The statement $x \in \mathrm{Cl}_t^{\succeq}$ means ''x belongs at least to class Cl_t'', while $x \in \mathrm{Cl}_t^{\preceq}$ means ''x belongs at most to class Cl_t''. Observe that $\mathrm{Cl}_1^{\succeq}=\mathrm{Cl}_n^{\preceq}=U; \mathrm{Cl}_n^{\succeq}=\mathrm{Cl}_n; \mathrm{Cl}_1^{\preceq}=\mathrm{Cl}_1$.

The key idea of the rough set philosophy is approximation of one knowledge by another knowledge. In our case, the knowledge being approximated is a collection of upward and downward unions of classes and the "granules of knowledge" are sets of objects defined using indiscernibility, similarity and outranking relations together.

Definition 15. With respect to $P \subseteq C$, the set of all objects belonging to $\mathrm{Cl}_t^{\succeq}$ without any left ambiguity constitutes the P^l-lower approximation of $\mathrm{Cl}_t^{\succeq}$, denoted by $\underline{P}^l(\mathrm{Cl}_t^{\succeq})$, and the set of all objects that could belong to $\mathrm{Cl}_t^{\succeq}$ constitutes the P^l-upper approximation of $\mathrm{Cl}_t^{\succeq}$, denoted by $\overline{P}^l(\mathrm{Cl}_t^{\succeq})$, for $t=1, \ldots, n$:

$$\underline{P}^l(\mathrm{Cl}_t^{\succeq})=\{x \in U: R_P^{l\succeq}(x) \subseteq \mathrm{Cl}_t^{\succeq}\},$$

$$\overline{P}^l(\mathrm{Cl}_t^{\succeq})=\{x \in U: R_P^{r\preceq}(x) \cap \mathrm{Cl}_t^{\succeq} \neq \varnothing\}.$$

Definition 16. The P^l-boundary of the unions $\mathrm{Cl}_t^{\succeq}$ are defined as

$$\mathrm{Bn}_P^l(\mathrm{Cl}_t^{\succeq})=\overline{P}^l(\mathrm{Cl}_t^{\succeq})-\underline{P}^l(\mathrm{Cl}_t^{\succeq}).$$

Definition 17. With respect to $P \subseteq C$, the set of all objects belonging to $\mathrm{Cl}_t^{\succeq}$ without any right ambiguity constitutes the P^r-lower approximation of $\mathrm{Cl}_t^{\succeq}$, denoted by $\underline{P}^r(\mathrm{Cl}_t^{\succeq})$, and the set of all objects that could belong to $\mathrm{Cl}_t^{\succeq}$ constitutes the P^r-upper approximation of $\mathrm{Cl}_t^{\succeq}$, denoted by $\overline{P}^r(\mathrm{Cl}_t^{\succeq})$, for $t=1, \ldots, n$:

$$\underline{P}^r(\mathrm{Cl}_t^{\succeq})=\{x \in U: R_P^{r\succeq}(x) \subseteq \mathrm{Cl}_t^{\succeq}\},$$

$$\overline{P}^r(\mathrm{Cl}_t^{\succeq})=\{x \in U: R_P^{l\preceq}(x) \cap \mathrm{Cl}_t^{\succeq} \neq \varnothing\}.$$

Definition 18. The P^r-boundary of the unions $\mathrm{Cl}_t^{\succeq}$ are defined as

$$\mathrm{Bn}_P^r(\mathrm{Cl}_t^{\succeq})=\overline{P}^r(\mathrm{Cl}_t^{\succeq})-\underline{P}^r(\mathrm{Cl}_t^{\succeq}).$$

Analogously, using $R_P^{l\preceq}$ and $R_P^{r\preceq}$, we can define P^l-lower approximation, P^l-upper approximation, and P^r-lower approximation, P^r-upper approximation of $\mathrm{Cl}_t^{\preceq}$, for $t=1, \ldots, n$:

$$\underline{P}^l(\mathrm{Cl}_t^{\preceq})=\{x \in U: R_P^{l\preceq}(x) \subseteq \mathrm{Cl}_t^{\preceq}\},$$

$$\overline{P}^l(\mathrm{Cl}_t^{\preceq})=\{x \in U: R_P^{r\succeq}(x) \cap \mathrm{Cl}_t^{\preceq} \neq \varnothing\}.$$

$$\underline{P}^r(\mathrm{Cl}_t^{\preceq})=\{x \in U: R_P^{r\preceq}(x) \subseteq \mathrm{Cl}_t^{\preceq}\},$$

$$\overline{P}^r(\mathrm{Cl}_t^{\preceq})=\{x \in U: R_P^{l\succeq}(x) \cap \mathrm{Cl}_t^{\preceq} \neq \varnothing\}.$$

Definition 19. The P^l-boundary and the P^r-boundary of the unions $\mathrm{Cl}_t^{\preceq}$ are defined, respectively, as

$$\mathrm{Bn}_P^l(\mathrm{Cl}_t^{\preceq})=\overline{P}^l(\mathrm{Cl}_t^{\preceq})-\underline{P}^l(\mathrm{Cl}_t^{\preceq}),$$

$$\mathrm{Bn}_P^r(\mathrm{Cl}_t^{\preceq})=\overline{P}^r(\mathrm{Cl}_t^{\preceq})-\underline{P}^r(\mathrm{Cl}_t^{\preceq}).$$

IV. PROPERTIES

Theorem 1. (Rough inclusion). For any $t \in T$ and for any $P \subseteq C$,

$$\underline{P}^{l}(\mathrm{Cl}_t^{\succeq}) \subseteq \mathrm{Cl}_t^{\succeq} \subseteq \overline{P}^{l}(\mathrm{Cl}_t^{\succeq}),$$
$$\underline{P}^{r}(\mathrm{Cl}_t^{\succeq}) \subseteq \mathrm{Cl}_t^{\succeq} \subseteq \overline{P}^{r}(\mathrm{Cl}_t^{\succeq}),$$
$$\underline{P}^{l}(\mathrm{Cl}_t^{\preceq}) \subseteq \mathrm{Cl}_t^{\preceq} \subseteq \overline{P}^{l}(\mathrm{Cl}_t^{\preceq}),$$
$$\underline{P}^{r}(\mathrm{Cl}_t^{\preceq}) \subseteq \mathrm{Cl}_t^{\preceq} \subseteq \overline{P}^{r}(\mathrm{Cl}_t^{\preceq}).$$

Proof.

Let $x \in \underline{P}^{l}(\mathrm{Cl}_t^{\succeq})$, then $R_P^{l\succeq}(x) \subseteq \mathrm{Cl}_t^{\succeq}$. Since $R_P^{l\succeq}$ being reflexive, we have $x \in R_P^{l\succeq}(x)$. So $x \in \mathrm{Cl}_t^{\succeq}$. Let $x \in \mathrm{Cl}_t^{\succeq}$. Since $x \in R_P^{l\succeq}(x)$, which implies $x \in \overline{P}^{l}(\mathrm{Cl}_t^{\succeq})$. The remaining proofs are analogous.

Theorem 2. For any $t \in T$ and for any $P \subseteq C$,

$$\overline{P}^{l}(\mathrm{Cl}_t^{\succeq}) = \bigcup_{x \in \mathrm{Cl}_t^{\succeq}} R_P^{l\succeq}(x),\ \overline{P}^{r}(\mathrm{Cl}_t^{\succeq}) = \bigcup_{x \in \mathrm{Cl}_t^{\succeq}} R_P^{r\succeq}(x),$$
$$\overline{P}^{l}(\mathrm{Cl}_t^{\preceq}) = \bigcup_{x \in \mathrm{Cl}_t^{\preceq}} R_P^{l\preceq}(x),\ \overline{P}^{r}(\mathrm{Cl}_t^{\preceq}) = \bigcup_{x \in \mathrm{Cl}_t^{\preceq}} R_P^{r\preceq}(x).$$

Proof.

$R_P^{r\preceq}(x) \cap \mathrm{Cl}_t^{\succeq} \neq \emptyset$ means that there is at least one $y \in \mathrm{Cl}_t^{\succeq}$ such that $y \in R_P^{r\preceq}(x)$, which implies $x \in R_P^{l\succeq}(y)$. Therefore $\overline{P}^{l}(\mathrm{Cl}_t^{\succeq})$ is composed of all $x \in R_P^{l\succeq}(y)$ where $y \in \mathrm{Cl}_t^{\succeq}$, that is, $\overline{P}^{l}(\mathrm{Cl}_t^{\succeq}) = \bigcup_{y \in \mathrm{Cl}_t^{\succeq}} R_P^{l\succeq}(y)$, which represents the result searched for. The remaining proofs are analogous.

Theorem 3. (Complementarity) . For any $t \in T$ and for any $P \subseteq C$,

$$\underline{P}^{l}(\mathrm{Cl}_t^{\succeq}) = U - \overline{P}^{r}(\mathrm{Cl}_{t-1}^{\preceq}),$$
$$\underline{P}^{r}(\mathrm{Cl}_t^{\succeq}) = U - \overline{P}^{l}(\mathrm{Cl}_{t-1}^{\preceq}),$$
$$\underline{P}^{l}(\mathrm{Cl}_t^{\preceq})\} = U - \overline{P}^{r}(\mathrm{Cl}_{t+1}^{\succeq}),$$
$$\underline{P}^{r}(\mathrm{Cl}_t^{\preceq}) = U - \overline{P}^{l}(\mathrm{Cl}_{t+1}^{\succeq}).$$

Proof.

If $x \in \underline{P}^{l}(\mathrm{Cl}_t^{\succeq})$, then $R_P^{l\succeq}(x) \subseteq \mathrm{Cl}_t^{\succeq}$, and, therefore, there is no $y \notin \mathrm{Cl}_t^{\succeq}$ such that $y\, R_P^{l\succeq} x$. Since the complement of $\mathrm{Cl}_t^{\succeq}$ in U is $\mathrm{Cl}_{t-1}^{\preceq}$, we can also say that $x \in \underline{P}^{l}(\mathrm{Cl}_t^{\succeq})$ if and only if there is no $y \in \mathrm{Cl}_{t-1}^{\preceq}$ such that $y\, R_P^{l\succeq} x$. Since $x \in \overline{P}^{r}(\mathrm{Cl}_{t-1}^{\preceq})$ if and only if there exists at least one $y \in U$ such that $y \in \mathrm{Cl}_{t-1}^{\preceq}$ and $y\, R_P^{l\succeq} x$, this means that $x \in \underline{P}^{l}(\mathrm{Cl}_t^{\succeq})$ if and only if $x \notin \overline{P}^{r}(\mathrm{Cl}_{t-1}^{\preceq})$. Thus, remembering the definitions of $\underline{P}^{l}(\mathrm{Cl}_t^{\succeq})$ and $\overline{P}^{r}(\mathrm{Cl}_{t-1}^{\preceq})$, we have proved that $\underline{P}^{l}(\mathrm{Cl}_t^{\succeq}) = U - \overline{P}^{r}(\mathrm{Cl}_{t-1}^{\preceq})$. The remaining proofs are analogous.

Theorem 4. (Identity of boundaries) . For any $t \in T - \{1\}$ and for any $P \subseteq C$,

$$\mathrm{Bn}_P^{l}(\mathrm{Cl}_t^{\succeq}) = \mathrm{Bn}_P^{r}(\mathrm{Cl}_{t-1}^{\preceq}),\ \mathrm{Bn}_P^{r}(\mathrm{Cl}_t^{\succeq}) = \mathrm{Bn}_P^{l}(\mathrm{Cl}_{t-1}^{\preceq}).$$

Proof.

From Theorem 3 we have

$$\mathrm{Bn}_P^{l}(\mathrm{Cl}_t^{\succeq}) = \overline{P}^{l}(\mathrm{Cl}_t^{\succeq}) - \underline{P}^{l}(\mathrm{Cl}_t^{\succeq}) = [U - \underline{P}^{r}(\mathrm{Cl}_{t-1}^{\preceq})] - [U - \overline{P}^{r}(\mathrm{Cl}_{t-1}^{\preceq})] = \overline{P}^{r}(\mathrm{Cl}_{t-1}^{\preceq}) - \underline{P}^{r}(\mathrm{Cl}_{t-1}^{\preceq}) = \mathrm{Bn}_P^{r}(\mathrm{Cl}_{t-1}^{\preceq}).$$

$$\mathrm{Bn}_P^{r}(\mathrm{Cl}_t^{\succeq}) = \overline{P}^{r}(\mathrm{Cl}_t^{\succeq}) - \underline{P}^{r}(\mathrm{Cl}_t^{\succeq}) = [U - \underline{P}^{l}(\mathrm{Cl}_{t-1}^{\preceq})] - [U - \overline{P}^{l}(\mathrm{Cl}_{t-1}^{\preceq})] = \overline{P}^{l}(\mathrm{Cl}_{t-1}^{\preceq}) - \underline{P}^{l}(\mathrm{Cl}_{t-1}^{\preceq}) = \mathrm{Bn}_P^{l}(\mathrm{Cl}_{t-1}^{\preceq}).$$

Theorem 5. (Monotonicity). For any $t \in T$ and for any $P \subseteq Q \subseteq C$,

$$\underline{P}^{l}(\mathrm{Cl}_t^{\succeq}) \subseteq \underline{Q}^{l}(\mathrm{Cl}_t^{\succeq}),\ \underline{P}^{l}(\mathrm{Cl}_t^{\preceq}) \subseteq \underline{Q}^{l}(\mathrm{Cl}_t^{\preceq}),$$
$$\overline{P}^{l}(\mathrm{Cl}_t^{\succeq}) \supseteq \overline{Q}^{l}(\mathrm{Cl}_t^{\succeq}),\ \overline{P}^{l}(\mathrm{Cl}_t^{\preceq}) \supseteq \overline{Q}^{l}(\mathrm{Cl}_t^{\preceq}),$$
$$\underline{P}^{r}(\mathrm{Cl}_t^{\succeq}) \subseteq \underline{Q}^{r}(\mathrm{Cl}_t^{\succeq}),\ \underline{P}^{r}(\mathrm{Cl}_t^{\preceq}) \subseteq \underline{Q}^{r}(\mathrm{Cl}_t^{\preceq}),$$
$$\overline{P}^{r}(\mathrm{Cl}_t^{\succeq}) \supseteq \overline{Q}^{r}(\mathrm{Cl}_t^{\succeq}),\ \overline{P}^{r}(\mathrm{Cl}_t^{\preceq}) \supseteq \overline{Q}^{r}(\mathrm{Cl}_t^{\preceq}),$$
$$\mathrm{Bn}_P^{l}(\mathrm{Cl}_t^{\succeq}) \supseteq \mathrm{Bn}_Q^{l}(\mathrm{Cl}_t^{\succeq}),\ \mathrm{Bn}_P^{l}(\mathrm{Cl}_t^{\preceq}) \supseteq \mathrm{Bn}_Q^{l}(\mathrm{Cl}_t^{\preceq}),$$
$$\mathrm{Bn}_P^{r}(\mathrm{Cl}_t^{\succeq}) \supseteq \mathrm{Bn}_Q^{r}(\mathrm{Cl}_t^{\succeq}),\ \mathrm{Bn}_P^{r}(\mathrm{Cl}_t^{\preceq}) \supseteq \mathrm{Bn}_Q^{r}(\mathrm{Cl}_t^{\preceq}).$$

Proof.

From the definition of the global binary relations it follows that for any $P \subseteq Q \subseteq C$, $x\, R_Q^{l\succeq} y$ implies $x\, R_P^{l\succeq} y$ and $x\, R_Q^{r\preceq} y$ implies $x\, R_P^{r\preceq} y$, for each $x, y \in U$. From the first implication we have $R_Q^{l\succeq}(x) \subseteq R_P^{l\succeq}(x)$ for each $x \in U$ and, in consequence, $R_P^{l\succeq}(x) \subseteq \mathrm{Cl}_t^{\succeq}$ implies $R_Q^{l\succeq}(x) \subseteq \mathrm{Cl}_t^{\succeq}$. From the second implication it follows that $R_Q^{r\preceq}(x) \subseteq R_P^{r\preceq}(x)$ for each $x \in U$. Thus, on the basis of the definition of approximations and boundary of $\mathrm{Cl}_t^{\succeq}$, we obtain the proof with respect to $\mathrm{Cl}_t^{\succeq}$. Analogous proof holds for $\mathrm{Cl}_t^{\preceq}$.

ACKNOWLEDGMENT

This paper is supported by the National Natural Science Foundation of China (No. 70601013).

REFERENCES

[1] Z. Pawlak, "Rough sets", International Journal of Computer and Information Sciences, Vol. 11, No. 5, pp. 341-356, 1982.

[1] Z. Pawlak, Rough Sets–Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht, Boston, London, 1991.

[2] Y. Y. Yao and J.P. Zhang, Interpreting fuzzy membership functions in the theory of rough sets, Lecture Notes In Computer Science; Vol. 2005, pp.82-89, Springer Berlin / Heidelberg, 2000.

[3] W. Zhu. "Generalized rough sets based on relations", Information Sciences, Vol. 177, No. 22, pp. 4997-5011, 2007.

[4] M. Kondo, "On the structure of generalized rough sets", Information Sciences, Vol. 176, No. 5, pp. 589-600, 2006.

[5] Y. Yao, On generalizing Pawlak approximation operators, in: Lecture Notes In Computer Science, Vol. 1424, pp. 298-307. Springer-Verlag, London, UK, 1998.

[6] Y. Yao, "Constructive and algebraic methods of theory of rough sets", Information Sciences, Vol. 109, No. 1-4, pp. 21–47, 1998.

[7] S. Greco, B. Matarazzo, and R. Slowinski, “Rough set theory for multicriteria decision analysis”, European Journal of Operational Research, Vol. 129, No. 1, pp. 1-47, 2001.

[8] S. Greco, B. Matarazzo, and R. Slowinski, “Rough sets methodology for sorting problems in presence of multiple attributes and criteria”, European Journal of Operational Research, Vol. 138, No. 2, pp. 247-259, 2002.

[9] W. Zhu, and F. Y. Wang. “Binary relation based rough sets”, Lecture Notes in Computer Science, Vol. 4223/2006, pp. 276-285, Springer Berlin, Heidelberg, 2006.

[10] Y. Y. Yao, “Relational interpretations of neighborhood operators and rough set approximation operators”, Information Sciences, Vol. 111, No. 1-4, pp. 239-259, 1998.

[11] R. Slowinski, and D. Vanderpooten, “Similarity relations as a basis for rough approximations,” in Advances in Machine Intelligence and Soft computing, Vol. IV, pp. 17-33, Durham, NC: Duke University Press, 1997.

[12] R. Slowinski, D. Vanderpooten, “A generalized definition of rough approximations based on similarity”, IEEE Transactions on Knowledge and Data Engineering, vol. 12, no. 2, pp. 331-336, 2000.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 049-051

The Research of Direct Routing Technology Implementing the Load Balancing

Yan Gao [1,2], Zhibin Zhang [2], and Weifeng Du [3*]
[1] Intelligent Control and Development Center, Southwest Jiaotong University, Chengdu, China
Email: gaoyan@hpu.edu.cn
[2] School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China
Email: zhangzhibin@hpu.edu.cn
[3] The School of Mathematics & Information Engineering, Jiaxing University, Jiaxing, China, 314001
Email:woodmud@tom.com

***Abstract*—When the network is using many servers to provide network services, it need a load balancing technique. This article has researched one of load balancing technique that is direct routing. Direct routing technology program is deployed and running on speed Ethernet or kilo-mega Ethernet; it needn't to change the original topology. The client sends the data request; the load-balancing device is only responsible for dispatching the request, while the application server will respond back to the client. Under this load balancing mode, the overall performance of the system does not rely on the performance of the device of load balancing performance itself, but it depends on the performance of the application server itself. It can maximize server performance.**

***Index Terms*—load balancing, direct routing, load Coordinator, Server pool**

I. INTRODUCTION

At present, no matter what in the enterprise network, campus network or WAN, the development of the portfolio of the Web services has exceeded the most optimistic estimates of the past. With the popularity and applications of Internet, even according to the optimal allocation to construct the network at that time, it is hard to meet the normal demand for services.

For example, a site will receive millions of times the access request each day, so hardware processing power will soon become a bottleneck for the server that provides large loads Web services. Simple increasing hardware performance can not really solve this problem, because a single server's performance is always limited. In particular, the network request usually occurs suddenly. When some major event occurs, the network access will dramatically increase, resulting in network bottlenecks.

In order to solve these problems, it must use many servers to provide network services and distribute network requests to these servers to share. When it distributes network request to a lot of servers, you need to use a load balancing technique. Through the load-balancing technology it may evenly distribute network requests to each server, and finally it settles the bottleneck problem. This article will research a load balancing of the direct routing technology; it can achieve good load balancing between servers.

II. LOAD EQUALIZATION COLONY

Load-balancing technology is the core of load equalization colony. When load equalization colony is running, normally it distributes the workload to a group back-end server through one or more front-end load balancer, thereby, it achieves the overall system high performance and high availability. This computer colony is sometimes referred to as server colony (Server Farm).

In general, load equalization colony uses two-tier structure, double major components:

A. Load Coordinator (Load Balancer)

It is the front-end of machines of the entire colony in the outside, being responsible for sending the client's request to a group of servers to be carried out, while the customers think the service is from an IP address (we may call the virtual IP address);

B. Server Pool (Server Pool)

It is a group of the servers that really implement the client request, including the services of WEB, MAIL, FTP, and DNS, etc;

The coordinator is the only entry point (Single Entry Point) of the server colony system. It can use IP load balancing technology, content-based request distribution technology or combination of both. In the IP load-balancing technology, it needs the server pool to have the same contents and provide the same services. When a client request arrives, the coordinator selects a server from the server pool only according to server load state and the stetted scheduling algorithm. It transmits the request to the selected server, and records this dispatch; when other message of the request reaches, it would be also transmitted to the selected servers of the front. In the distribution technology that is based on content request, the server can provide different services. When customers request arrives, the coordinator may select the server to execute the request according to the request contents. Because all operations will be completed in the kernel space of the Linux operating system, and its dispatch cost is small, it has a very high throughput.

* Corresponding author: Du Weifeng , School of Mathematics & Information Engineering, Jiaxing University, Jiaxing, Zhejiang, China, Email:woodmud@tom.com

AP-PROC-CS-09CN005

The number of server pool nodes is variable. When the load that the whole system receives exceeds the current handling capacity of all nodes, you can increase the server in the server pool to meet the growing requests load. For most Internet services, the strong correlation between the requests does not exist; the request may be executed on different nodes in parallel, so the performance of the whole system increases linearly basically with the number of nodes of servers pool increasing.

III. Load Equalization Colony Performance Requirements

In the design, it need to consider the entire system of high-performance, high availability, scalability, manageability and high security. Colony system characteristic is that it has redundancy on both its hardware and software. The system's high availability is achieved through detecting node or service process fault and correctly resetting the system. Load equalization colony's high availability is reflected in two aspects:

A. Self-System High Availability

Coordinator itself can build high-availability pairs of machine system. When the host is system failure, spare machine turns into the host. When the original host is restored, it automatically switches back to the original state. At present, Front-end co-coordinator may become a single failure point of system (Single Point of Failure). In general, coordinator itself has the higher reliability because the fewer process running on coordinator and core processes has been traversed long before, but we can not rule out the major fault such as the aging hardware, network lines and human misuse and so on. In order to avoid coordinator failure to result in dysfunctions of the entire system, we need to set up subordinate coordinator as the backup of the primary coordinator. Here we use a virtual coordinator Redundancy Protocol (VRRP: Virtual Router Redundancy Protocol) to build two high availability colonies of two coordinators. Virtual coordinator Redundancy Protocol (VRRP) is an option agreement; it can assign dynamically virtual coordinator responsibilities to the one of the LAN VRRP coordinator. The VRRP coordinator that controls IP address of the virtual coordinator is called as the primary coordinator, and it is responsible for transmitting data packets to those virtual IP addresses. Once the main coordinator is not available, this selection process has provided dynamic fault transfer mechanism, which allows the virtual IP address of coordinator can be used as the default first-hop coordinator of the terminal host.

B. High Availability of the Server Colony that are Managed by the Load Balancing Coordinator

Usually, we have monitoring process of the resources to monitor the health status of the server nodes on the coordinator at all times. When the server doesn't reach for the ICMP ping or detecting its network services does not respond in the specified time, the monitoring process of the resource would inform the operating system kernel to remove or invalidate the server from the scheduler list. In this way, the new service request will not be dispatched to bad nodes. Monitoring process of the resource reports the failure to the administrator by e-mail or pager. Once the monitoring processes detect the server coming back to work, it notifies the coordinator to add the server to scheduling list to dispatch. In addition, through the system management process, administrators can send commands to add the new machine ready to join the service to improve the processing performance of the system. It can also be cut out existing server from services, in order to carry out system maintenance for server. It may carry out the application server's "hot pull in-pull out".

IV. Direct Routing (Direct Routing) Technology Implementing Load Balancing

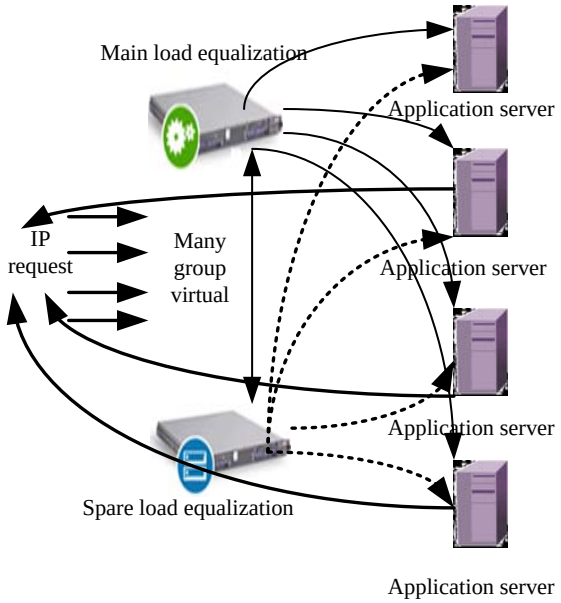


Figure 1. Load equalization of direct route technology.

Direct routing technology program was deployed and was running on speed Ethernet or kilo-mega Ethernet; it needn't to change the original topology. The client sends the data request; the load-balancing device is only responsible for dispatching the request, while the application server will respond back to the client.

In IP-based load-scheduling technology, when an initial SYN packet of TCP connection arrives, the coordinator selects an application server and transmits message to it. After this it ensures follow-up packets is transmitted to the server through searching IP and TCP packet header address. For the UDP data packet scheduling, the coordinator also will build schedule record and set the timeout value; in the set period, data packets that come from the same address and has the same service request (IP address and port) was dispatched to the same application server. After application server responds to customer's request, the data is no longer

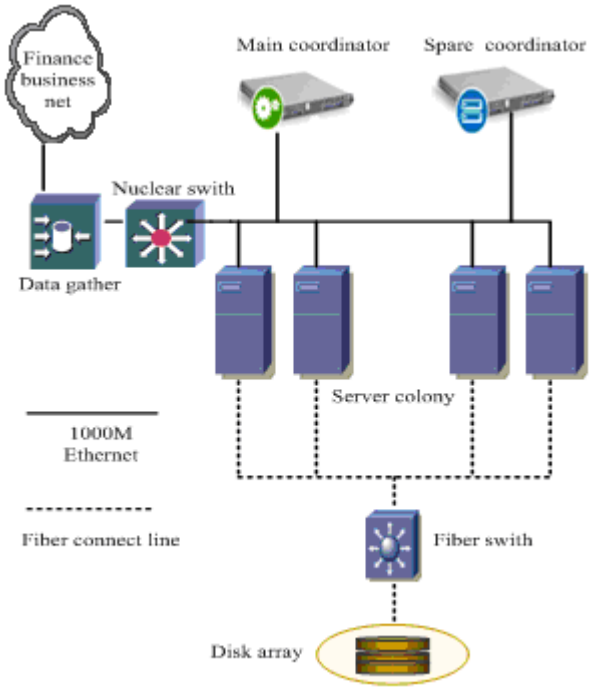


Figure 2. Testing Environment.

handled through the load-balancing device, but it directly returns to the client.

Under this load balancing mode, the overall performance of the system does not rely on the performance of the device of load balancing performance itself, but it depends on the performance of the application server itself. It can maximize server performance.

V. System Testing

In test environment, there is a large server that assumes data acquisition and preprocessing of services network. After the data obtained is pre-processed, it is sent to the four servers of the background analysis of the behavior through the load balancing function of the server of the main coordinator. The storage system is the SAN manner,

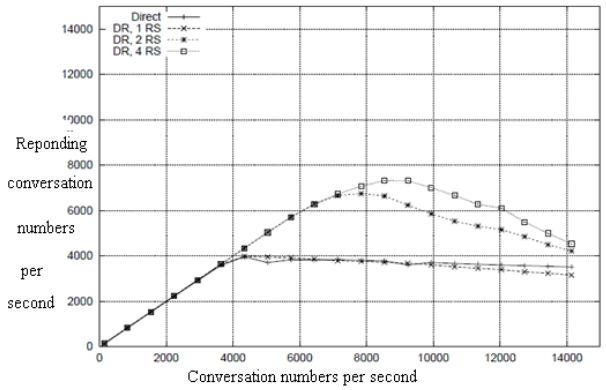


Figure 3. Direct route of load equalization performance.

a fiber switch and a large disk array. Two mutual backup co-coordinator complete high-availability of the load balancing system, and high availability of the server colony are implemented by the management function of the load balancing colony.

As shown above, we use direct routing (DR) technology to do load balancing respectively for 1-4 servers. When per second conversation exceed 8,000, the entire system bottlenecks will appear. But the increase of the number of servers has an evident effect on the improvement of the entire system performance.

VI. Conclusion

We can draw the following conclusions through testing:

Using the server load balancing of direct routing technology has the following advantages: it has superior performance with high network throughput; system construction cost is low; construction cost of load-balancing equipment itself is low; IP packets doesn't pass load balancing device; it improves the speed etc.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China (Grant No. 60875034), Henan province Natural Science Foundation (Grant No. 0611055800), Henan Province key attack project 082102210079, Henan province Science and Technology Tackle key Project (0424460013). This work was partially supported by Zhejiang province fatal project (priority subjects) key industrial project (2008C11011).

References

[1] Baltagi B H. Econometric analysis of panel data(third edition)New York Wiley.2005.

[2] Hsiao C. Analysis of panel data(second edition).Cambridge Cambridge University Press.

[3] Cemmeno R, Grier K.Conditional Hetemskedasticity and Cross-Sectional Dependence ineds. Contributions in Economic Analysis. Amsterdam Amsterdam Elsevier Press. 2006.

[4] Depreciation: Panel evidence for the G7 and 8 Latin American Countries. April. 2005.

[5] MUKHERJEE B.Optical communication networks[M] New York:Mc-Graw-Hill,1997.

[6] MURAKAMI M,MATSUDA T. Long-haul WDM transmission using higher order fiber dispersion management[J]. Lightwave Technology,2000,18(9):1197-1204.

[7] ARMITAGE J.CROCHAT O.LE B J-Y. Design of a survivable WDM photonic network[C]//Proceeding of IEEE INFOCOM'97.[S.I.]:IEEE Press,1997.

[8] NARVAEZ P,SIU K Y. Efficient algorithms for multi-path link state routing[C]//ISCOM'99[S.I]:IEEE Press,1999.

[9] GOJMERAC I, ZIEGLER T, REICHEL P.Adaptive multipath routing based on local distribution of link load information[C]//In Proceedings of QofIS. Stockholm: [s.n.],2003

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 052-057

Fault Tree Based Architectural Analysis for E-Business Systems

Wang Chu, and Yanli Feng
Shandong Institute of Business and Technology, Yantai, China
Email: sdchuw @ 163.com

***Abstract*—The increasing complexity of the e-business systems urges the improvement of existing methods of system analysis in order to reduce the likelihood that important threats remain unidentified. Such an improvement can be achieved by combining risk analysis methodologies with the architecture centric system design process. Existing methods are inherently deficient as far as vulnerability analysis for system architecture is concerned. This paper integrates Fault-Tree Analysis (FTA) technology and architecture centric system analysis method to analyze e-business system architecture. The analysis process sets focus on FTA driven scenarios generation and vulnerability analysis. The fault tree based architectural analysis approach is strongly architecture-centric. It can be used to discover component-level vulnerabilities. It is simple because the FTA diagram is intuitive and easy to be constructed. It provides an effective approach to e-business system architectural analysis.**

***Index Terms*—E-business, Fault tree analysis, E-business system architectural analysis, Quality attributes.**

I. INTRODUCTION

Internet technology is fast becoming a necessary component to building a competitive and successful business in today's "connected" economy. The term e-business is defined as "a secure, flexible, and integrated approach to delivering differentiated business value by combining the systems and processes that run core business operations with the simplicity and reach made possible by Internet technology". Business that can harness the power of the Internet by translating its competencies into value for its customer will enjoy a competitive advantage.

The dramatic growth of the applications deployed in World Wide Web have lead to the situation where the quality attributes have become very important concerns in the development of e-business systems. The key to success for e-business is ensuring data integrity, guaranteeing service availability and protecting confidentiality along the length of the entire online e-service chain [1].

The increasing complexity of the e-business systems urges the improvement of existing methods of system analysis in order to reduce the likelihood that important threats remain unidentified. Such an improvement can be achieved by combining risk analysis methodologies with respect to the system architecture. Determining the vulnerabilities embedded in e-business software systems and networks shall be the first step, which is very critical in practice [2,3]. System vulnerability analysis is concerned with ensuring and certifying that a delivered system does not pose an unacceptable danger to its end-users or to the environment in which the system is installed.

Many e-business system analysis approaches are currently being used in organizations. Existing methods are inherently deficient as far as vulnerability analysis for system architecture is concerned. There are three fundamental problems as following:

(1) E-business system analysis focuses on business perspectives not on IT perspective.

For e-business system analysis, no full-blown method currently exists. Most of the work concentrates on the strategic issues and choices. It is important to note that business model defines e-business application from the business perspective rather than the IT perspective [4, 5, 6].

(2) Quality goals are too abstract not concrete.

Current business analysis methodologies are inadequate because they are at a too high level and only address portions of the complete business analysis process [7]. Some vulnerabilities should be identified over component level such as the user, host, server, network, database, and so on.

(3) E-business system analysis is not architecture-centric.

The e-buiness system architecture is a key business asset for an organization, architectures are complex and involve many design tradeoffs, then architectural analysis must also be a key practice for that organization. Many vulnerabilities are caused not by individual component but by interactions between components, all the relevant components need to be identified from the system architecture design.

In this paper, we only focus on quality validation rather than on initial quality requirements formulation duo to the limited space, other processes of the architecture centric system analysis are not discussed in this paper. Our approach applies Fault-Tree Analysis technology to analysis of the e-business system architecture from IT perspective. The architecture analysis process focuses on FTA driven scenarios generation and vulnerabilities analysis.

The contributions of this paper consist in two aspects: 1) It supports top-down FTA driven scenarios generation and component vulnerability analysis at different abstraction-level. It is simple because the FTA diagram is intuitive and easy to be constructed; 2) It is strongly

AP-PROC-CS-09CN005

architecture-centric that is consistent with the architecture centric e-business system design method.

The rest of the paper is organized as follows. Section II presents some related work on e-business analysis and architecture analysis. Section III discusses how to analyze e-business system architecture based on FTA technology. Section IV presents a case study. Section V contains concluding remarks and future work.

II. Related Work

T. Dimitrakos et al. discuss a tool-supported framework for precise, unambiguous, and efficient risk assessment of security critical systems, its application on Web-enabled B2C e-commerce services and the meta-data based deployment model. The risk assessment process consists of steps: identify context, identify risk, analyse risks, evaluate risks, and treat risks [2].

Yudistira Asnar and Paolo Giorgini propose a framework to support modeling and analysis of business continuity plans from the organization perspective, where risks and treatments are modeled and analyzed along strategic objectives and their realizations. The goal-risk framework is used to analyze risk and to capture interdependencies among assets. A business continuity plan specifies the methodologies and procedures required to backup and recover every functional units of the business [8].

Stilianos Vidalis and Andy Jones present vulnerability trees technique based on the Object Oriented principles, the utility theory, FTAs and other methodologies by which the users identify key vulnerabilities that are common to more than one assets of the system and help them to counter them in a cost effective manner. The technique is part of a process that is aiming in minimising and controlling the threats against e-business [9].

Jaap Gordijn and Hans Akkermans present a conceptual modeling approach to e-business that is designed to help define how economic value is created and exchanged within a network of actors. The evaluation approach consists of creating profit sheets, evaluating the objects in the profit sheet in terms of their cost and benefit to the participating actors, and evaluating what-if scenarios. Analyzing what-if scenarios can also help find the weak and strong points of e-business models [10].

The Architecture Tradeoff Analysis Method (ATAM) is an analysis method that is used to assess the consequences of architectural decisions in light of quality attribute requirements. A prerequisite of an evaluation is to have a statement of quality attribute requirements and a specification of the architecture with a clear articulation of the architectural design decisions. Given the attribute requirements and the design decisions, the major goal of ATAM is to evaluate the architectural design decisions is to determine if they satisfactorily address the quality requirements [11].

III. Architectural Analysis for E-Business System

A. Architecture Centric System Design for E-business Application

In [12], we propose a component oriented e-business development approach that consists of four phases: Value-chain modeling, E-business design, E-service modeling, and E-business implementation. Components are classified into three types: business components, service components, and software components. The business components describe business entities related to reality world; the service components provide specific services to outside via interface and support implementation of business goal; and software components are usually regarded as executable components. In this solution, architecture is used as the blueprint for component composition, and component relationships are used to shorten the gap between business goal and services.

The component oriented development approach constructs a directed graph, in which the nodes are components and the edges reflect relationships in the direction of refinement. The bottom components in the graph are operationalised concrete components. The business processes can be depicted by component collaboration diagram and component sequence diagrams. The granularity of components for e-business is depicted in Figure 1. Figure 2 depicts the component diagram of an e-business system architecture.

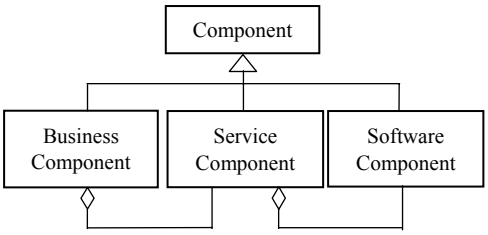


Fig.1 Component granularity

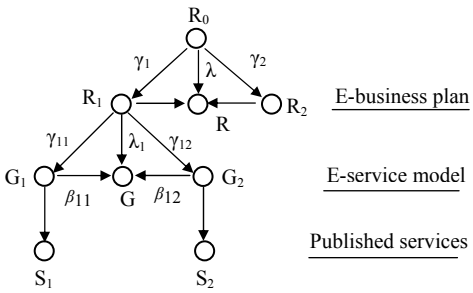


Fig.2 An e-business application

In Figure 2, γ-relationship is "Responsibility-Assignment" relationship between the system goal and the components. β-relationship is "Take-Part-In" relationship between the component and the architecture. λ-relationship is "Achieved-By" relationship between the system goal and the architecture.

The quality attributes should be "designed in" to an e-business system under development, rather than waiting

until the system is implemented and conducting a costly "test and rework". Without undertaking a formal analysis process, the organization cannot ensure that the architectural decisions made—particularly those which affect the achievement of quality attribute such as performance, availability, security, and modifiability—are advisable ones that appropriately mitigate risks. Hence, when E-service modeling phase is finishend the architectural analysis should be conducted.

B. Architecture-Centric System Analysis

The architecture-centric system analysis is meant to be a risk identification method, a means of detecting areas of potential risk within the architecture of a complex software intensive system. This has several implications:

- The risk identification can be done in the early phase of system development life cycle.
- It can be done inexpensively and quickly.
- The architectural analysis will produce analysis results commensurate with the component-level detail of the architectural specification.

The steps of the analysis process are as follows:

1) Generate concrete quality attribute scenarios. The quality factors that comprise system non-functional requirements (performance, availability, security, modifiability, etc.) are elicited, specified down to the concrete component-level scenarios that are prioritized.

2) Analyze architectural specification. Based upon the high-priority scenarios identified in Step 1, the architectural design that address those scenarios are analyzed.

3) Present analysis result. The uncovered issues are documented. The system architects will revise the architecture design.

C. Quality Attributes for E-business System

An e-business system has following quality attributes: security, modifiability, performance, and availability. Security is central to the success of the system since ensuring the privacy of the customers' data is of utmost importance; and modifiability is also essential to the success of system since we need to be able to respond quickly to a rapidly evolving and very competitive marketplace. Ensuring the availability of the key components is critical to online commercial success.

The utility tree is used to translate the business context first into quality attribute drivers and then into concrete scenarios that represent each business driver. The quality requirements are expressed as quality attribute scenarios for the system. These scenarios are prioritized in terms of how important they are to the overall mission of the system and the perceived risk in realizing them in the system. The highest priority scenarios will be used in the analysis of the architecture.

The system architectures are key to realizing quality attribute requirements. Although an architecture cannot guarantee that an implementation will meet its quality attribute goals, the wrong architecture will surely spell disaster. Components as well as communication mechanisms and paths must be designed or selected early in the life cycle to satisfy quality requirements.

In this paper, we set focus on architecture analysis, not on the quality requirements elicitation, therefore when analyzing architecture we assume that the quality requirements have been unambiguously specified.

D. FTA Based Architecture Analysis

A fault tree shows logical relationship between an event (failure) and its causes and provides a logical framework for expressing combinations of component failures that can lead to system failure. Fault tree analysis can be used to support engineering and management decisions, trade-off analysis and risk assessment.

A vulnerability is a weakness with respect to an asset or group of assets which can be exploited by one or more threats. There are six types of vulnerabilities that can exist in any system, and these are: Physical, Natural, Hardware/Software, Media, Communication, and Human. We need a process that will be able to analyze these different types [9].

(1) FTA driven scenarios generation

Given quality goal "g_1", "$\neg g_1$"expresses a failure that is used to construct a context specific fault tree. The fault tree diagram is used to identify the unusual, but possible combinations of component failures. This method is (also called flaw hypothesis) based on propagation of errors due to vulnerabilities in order to find possible ways to affect the e-business system. The idea is to hypothesize possible flaws, and then check whether these hypotheses are true. Figure 3 illustrates an example of fault tree.

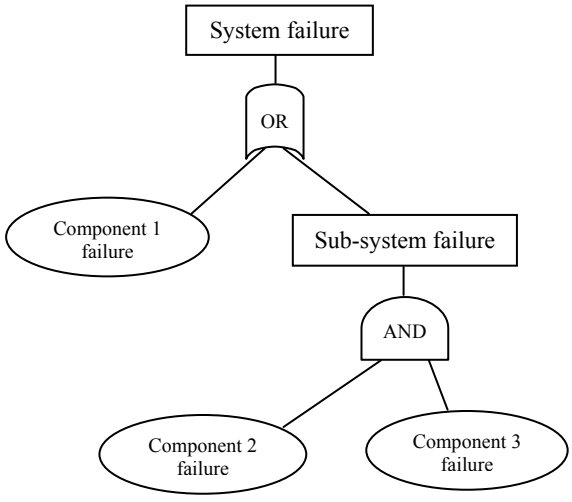


Fig. 3 An example of fault tree

A fault tree provides a threat model and is very useful to perform quality attributes analysis and evaluation. The fault tree analysis techniques start with the generation of cutsets. A cutset is a set of basic events, in this paper, we construct vulnerability scenarios based on the cutsets and the component interactions (sequence diagrams); if all the basic events in a cutset occur, then the top event (flaw hypothesis) occurs.

For each vulnerability consideration (flaw hypothesis), we identify a list of cutsets. This process of vulnerability

identification largely relies on the judgement and experience of the evaluators involved in the process.

(2) Vulnerability analysis

Fault-tree analysis is concerned with discovering the system states which are potentially vulnerable. Vulnerability analysis use the vulnerability scenarios to check different component connections of the e-business system architecture. Common vulnerabilities can be identified and countered in order to secure the system in a cost effective manner. Furthermore, critical vulnerability scenarios will differentiate the vulnerabilities that must be countered from those that will have to be countered some time in the future. Figure 4 shows the vulnerability analysis.

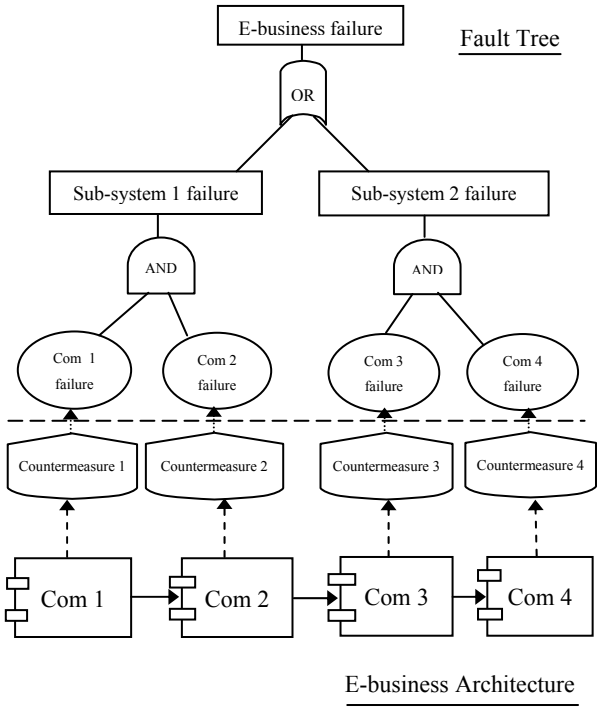


Fig. 4 Vulnerability analysis

Vulnerability analysis is conducted as following steps:

Step 1: For each quality feature (corresponding to a set of vulnerability scenarios), derive a countermeasures design model based on e-business system architecture specification.

Step 2: For each vulnerability scenario, analyze countermeasures for vulnerabilities.

The architect begins the process of mapping the scenarios onto whatever architectural descriptions have been presented. For example, some problems may be checked such as:

• What facilities exist in the system architecture for self-testing and monitoring of software components?

• What facilities exist in the system architecture for redundancy, liveness monitoring, failover, and how data consistency is maintained?

• What is the task view of the system, including mapping of these processes/tasks to hardware and the communication mechanisms between them?

• What functional dependencies exist among the system components?

• What data is kept in the database, how big is it, how much does it change, and who reads/writes it?

• What is the anticipated frequency and volume of data being transmitted among the system components?, and so on.

Step 3: Group the analysis results into: proper countermeasures, ineffective countermeasures, conflicting countermeasures, and omitted countermeasures.

The output of architectural analysis is used by designers to revise the e-business system architecture.

IV. CASE STUDY

Assume that we have design an architecture for an e-business system, Figure 5 depicts this architecture and Figure 6 shows a sequence diagram for a business service.

For example, the following scenario sheds light on the performance aspect of the quality requirement: *g*: "*A remote user requests a business service via the Web during peak usage and receives the response within four seconds.*" We will analyze whether the specified system architecture satisfies this performance attribute.

(1) FTA driven scenarios generation

At first, we construct fault tree based on ┐ *g*, showed as Figure 7.The cutsets are:{Channel 1 jam},{Web server run slowly}, {Channel 2 jam}, {Database server run slowly}.

Due to the limited space, we only consider the hardware platform (i.e., Server, Network), other factors are not discussed, such as software component deign, business process deign, and so on.

Secondly, vulnerability scenarios are identified based on the cutsets and component sequence diagram as follows:

Scenario 1: User formulate service request, Send request (>4s) [Channel 1 jam].

Scenario 2: User formulate service request, Send request, Web server preprocess request (>4s) [Web server run slowly].

Scenario 3: User formulate service request, Send request, Web server preprocess request, Database server process request (>4s) [Database server run slowly].

Scenario 4: User formulate service request, Send request, Web server preprocess request, Database server process request, Database server return result (>4s) [Channel 2 jam].

(2) Vulnerability analysis

Provided that designers have taken following measures for the e-business perpormance: *broadband access network*, *web server cluster*, *high-performance database server platform*, and these measures are described in architecture design specification.

For every vulnerability scenario, the evaluators check whether there exists countermeasure(s) and whether the countermeasure(s) satisfies performance requirement.

Finaly, the analysis results are grouped into proper countermeasures{*broadband access network*, *web server cluster*, *high-performance database server platform*} and omitted countermeasures{*link bandwidth between web server and database server*}.

Designers sholud revise their architecture design according to aforementioned analysis results.

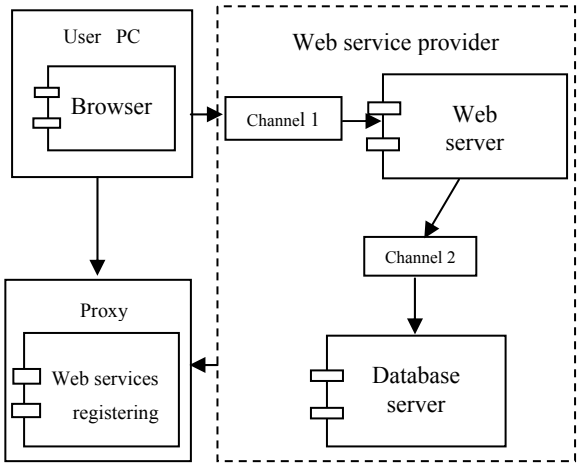


Fig.5 E-business system architecture

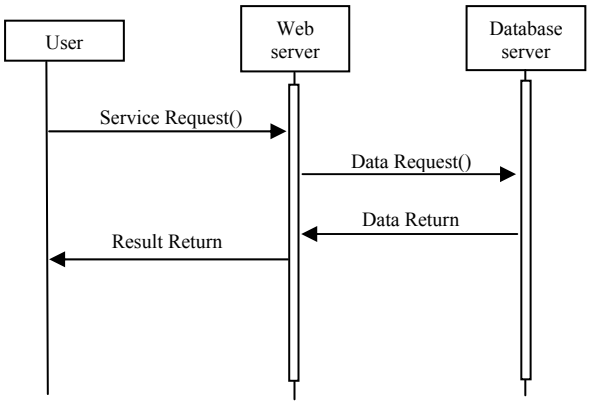


Fig. 6 A sequence diagram for the business service

V. Conclusion and Future Work

The main characteristics of the fault tree based architectural analysis approach for e-business systems are listed as following: Firstly, it combines FTA technology with the architecture-centric e-business design method seamlessly. The architecture-centric design method has been widely used in computer intensive system design, from IT perspective, e-business systems are also computer intensive systems. Hence our approach is suitable to e-business system design. Secondly, the fault tree based architectural analysis is strongly architecture-centric. It can be used to discover component-level vulnerabilities. Thirdly, the fault tree based architectural analysis is simple and powerful. It is simple because the FTA diagram is intuitive and easy to be constructed. It is powerful because the the fault tree based architectural analysis approach provides an effective process for vulnerability analysis of e-business system architecture.

The future work will concentrate on following aspects: automated generation of the vulnerability scenarios and visual analysis tool.

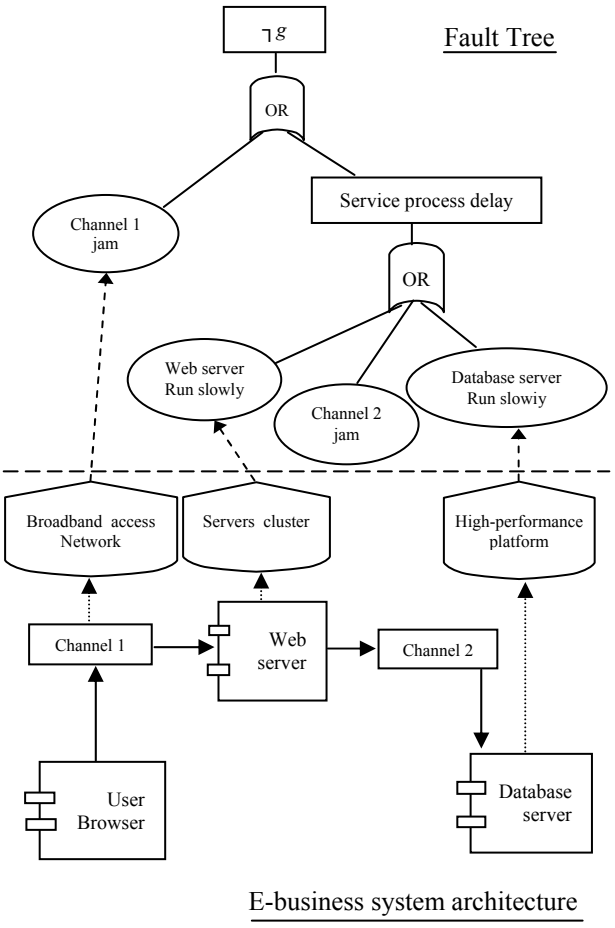


Fig.7 Vulnerability analysis

References

[1] "Secure E-business", Technical Report, Internet Security Systems, 2000.

[2] T. Dimitrakos, B. Ritchie, D. Raptis, K. Stølen. "Model based Security Risk Analysis for Web Applications: The CORAS approach". *Euroweb 2002 — The Web and the GRID: from e-science to e-business*. pp.1-13.

[3] K. Chandra Sekaran. "Requirements Driven Multiple View Paradigm for Developing Security Architecture". World Academy of Science, *Engineering and Technology* 31 2007. pp. 156-159.

[4] Maria-Eugenia Iacob, Piet Boekhoudt, Erwin Fielt. "E-Business Analysis Handbook", Telematica Instituut, 2001.

[5] Wil Janssen, Maarten W. A. Steen, and Henry Franken. "Business Process Engineering versus E-Business

Engineering". *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03) - Track 7*, Vol. 7, 2003. pp.185c.

[6] N. Nayak,M. Linehan, A. Nigam, et al. "Core business architecture for a service-oriented enterprise". *IBM Systems Journal*, Vol. 46, No. 4, 2007. pp.723-742.

[7] Avsharn Bachoo. "A Business Analysis Methodology". Thesis, University of Witwatersrand. 2006.

[8] Yudistira Asnar, Paolo Giorgini. "Analyzing Business Continuity through a Multi-Layers Model". *Proceedings of the 6th International Conference on Business Process Management*, 2008. pp.212 – 227.

[9] Stilianos Vidalis and Andy Jones. "Using Vulnerability Trees for Decision Making in Threat Assessment". Technical Report, CS-03-2, University of Glamorgan, June 2003.

[10] Jaap Gordijn and Hans Akkermans. "Designing and Evaluating E-Business Models". *IEEE Intelligent Systems*, July/August 2001. pp.11-17.

[11] Rick Kazman, Mark Klein, Paul Clements. "ATAM: Method for Architecture Evaluation". August 2000, Technical Report, CMU/SEI-2000-TR-004.

[12] Wang Chu, Depei Qian. "A Component-Oriented Development Approach to E-Business Applications". *Proceedings of IEEE International Conference on e-Business Engineering.* October 22-24, 2008, Xi'an China. IEEE Press. pp.45-52.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 058-061

Detailed Soot Source Terms Modeling in Turbulent Reacting Flow

Yongfeng Liu [1,2] , Youtong Zhang [1] , Hongsen Tian[2] , and Lianda Liu[1]
1. School of Mechanical and Vehicular Engineering, Beijing Institute of Technology, Beijing 100081, China
2. School of Mechanical and Electronic and Automobile Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China
Email:liuyongfeng@bucea.edu.cn;youtong@bit.edu.cn

***Abstract*—To calculate soot source terms a new detailed kinetic soot model is applied to study the formation and oxidation of soot particles in turbulent flames. The model is based on a detailed description of the chemical and physical processes leading to the formation of soot. It can be subdivided into the growth of polycyclic aromatic hydrocarbons (PAHs) in the gas phase reactions and the processes of particle inception, condensation, surface growth, and oxidation. Two different parts are developed about the growth of PAHs in the gas phase reaction. The first step towards the formation of soot is the formation of benzene and the second step is how to form PAH from benzene. In surface growth and oxidation process Hydrogen Abstraction - Carbon Addition (HACA) mechanism is modified due to the finding that the bound between the acetylene and the soot surface can be broken at high temperature in the experiment. Finally the analysis about soot source terms are used in turbulent combustion and different calculated results are obtained for different soot source terms. Acetylene, soot and OH densities are discussed and soot spatial distributions in the cylinder for different crank angle degree are carried out in 4JB1 engine.**

***Index Terms*—Soot source terms; Polycyclic aromatic hydrocarbons ;(PAHs), Turbulent combustion1 Introduction**

I. Introduction

Calculation of soot forming and oxidation in diesel engines can mainly be classified in different model approaches [1] -[2]. Balthasar[3] et al focused mainly on the fluid dynamics using simple models for the chemical processes. Soot formation is modeled with the help of empirical or semi-empirical models. The most recent approach, is based on the laminar flamelet concept [4]-[5], solving so-called Representative Interactive Flamelets (RIF) on line to the CFD code.

The concept in the present study is to use different approaches to get soot source terms. In gas phase a simple model is used to describe the growth of PAHs. Two parts are divided to understand the growth of PAHs easily. In particle inception and condensation stages Smoluchowski's equation is developed in terms of moments of the soot particle size distribution. In surface growth and oxidation stages HACA-mechanism is changed to a new mechanism. This is due to the finding that the bound between the acetylene and the soot surface can be broken at high temperature. The source terms due to particle coagulation has been omitted, since this process does not contribute to the growth in soot volume fraction, and is thus equal to zero at all times. The rates of soot surface growth and oxidation are normalized by the local soot volume fraction to account for the surface dependence of these processes. In the present study the rates of soot formation are in contrast to other approaches to model soot formation in turbulent reacting flow calculated on the basis of a detailed kinetic soot model.

II. Soot Formation and Oxidation

A. The Growth of PAHs

By multiplying these equations by the number of monomer units to the power of *r* one obtains the equation for the *r*-th moment of the PAH size distribution. The moments are defined as:

$$\left[M_r^{PAH}\right] = \sum_{i=1}^{\infty}\sum_{j=1}^{6} n_{i,j}^{r}[P_{i,j}] \quad r = 0,1,\ldots \propto \quad . \tag{1}$$

Where $[P_{i,j}]$ is the concentration of the PAH and *j* is in a step *i*, *n* is the monomer unit which is one C-Atom, and *r* is the number of the moment.

B. Particle Inception

Particle inception can be modeled as the coagulation of two PAH-molecules. Coagulation of particles of the same type can be described by Smoluchowski's equation [6]:

$$\dot{N}_i = \frac{1}{2}\sum_{j=1}^{i-1}(\beta_{j,i-j}N_j N_{i-j}) - \sum_{j=1}^{\infty}(\beta_{i,j}N_i N_j) \tag{2}$$

This equation gives the change of the particle number in size-class *i* as a function of time. The first terms in eq. (2) describes the formation of new particles from smaller sized particles and the second terms the consumption of particles in the *i*-th size-class by coagulation with particles of all size classes. The frequency factor $\beta_{i,j}$ in the free molecular regime is given by:

$$\beta_{i,j} = \varepsilon_{i,j}\sqrt{\frac{8\pi k_B T}{\mu_{i,j}}}\left(r_i + r_j\right)^2 . \tag{3}$$

Where k_B is the Boitzmann constant, $\mu_{i,j}$ is the reduced mass, r_i is the radius of particles of class i and $\varepsilon_{i,j}$ is the size dependent coagulation enhancement factor due to attractive or repulsive forces between the particles. The Smoluchowiski equation can be formulated for the particle inception omitting the second terms:

$$\dot{N}_i = \frac{1}{2}\sum_{j=1}^{i-1}(\beta_{j,i-j} N^p{}_j N^p{}_{i-j}) \ . \quad (4)$$

Eq. (4) can by multiplying with i^r and a summation over all size classes is described in terms of the moments of the PAH size distribution (note the change of the upper summation limit with respect to *j*):

$$\dot{M}_{r,pi} = \frac{1}{2}\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left((i+j)^r \beta_{i,j} N^p{}_i N^p{}_j\right) \ . \quad (5)$$

One can rewrite (3) with the help of $V_i = m_i/\rho_s = i \cdot m_1/\rho_s$:

$$\beta_{i,j} = C \cdot \left(\frac{1}{i}+\frac{1}{j}\right)^{1/2}\left(i^{1/3}+j^{1/3}\right)^2 \quad (6)$$

with

$$C = \varepsilon_{i,j}\left(\frac{3m_1}{4\pi\rho_s}\right)^{1/6}\left(\frac{6k_BT}{\rho_s}\right)^{1/2} \ . \quad (7)$$

Where V_i and m_i are the volume and mass of particle of size class i respectively, ρ_s=1860 [kg/m^3] the density of soot and $\varepsilon_{i,j}$ the coagulation enhancement due to inter-particle forces. Based on results of Kennedy (1988), the enhancement factor due to van der Waals interaction is set to a constant value of 2.2 for particle inception, condensation:

$$C_{nn} = 2.2\left(\frac{3m_1}{4\pi\rho_s}\right)^{1/6}\left(\frac{6k_BT}{\rho_s}\right)^{1/2} \ . \quad (8)$$

where the subscription *nn* stands for neutral-neutral coagulation. For the particle inception it is assumed that the coagulating particles are of the same size (i=j). One obtains under this assumption $\beta_{i,j} = \alpha\ C_i^{1/6}$, with $\alpha = 4\sqrt{2}$. The source terms for the moments in respect to particle inception can now with (5) be formulated as function of the moments for the soot particle and PAH size distribution, denoted by *P*:

$$\dot{M}_{r,pi} = \frac{1}{2}C_{nn}\alpha\sum_{k=0}^{r}\binom{r}{k}\left(M^p{}_{k+\frac{1}{6}} M^p{}_{r-k}\right) \quad (9)$$

The moments $M_0 - M_r$ are obtained from the fast polymerization model for the PAH growth. The fractional-order moments have to be interpolated by a logarithmic Lagrange interpolation or extrapolation.

C. Condensation

Condensation modeled as the coagulation of PAH molecules with soot particles are also described by the Smoluchowski's equation (2). The Smoluchowski's equation for condensation has the following form:

$$\dot{N}_{i,con} = \sum_{j=1}^{i-1}(\beta_{j,i-j} N^p{}_j N^s{}_{i-j}) - \sum_{j=1}^{\infty}(\beta_{i,j} N^s{}_i N^p{}_j) \quad (10)$$

Where N_i^s is the number of soot particles in the size class i and $N^p{}_j$ is the number of PAHs of size class j. The factor ½ in front of the first terms is omitted since the particles are not of the same type. This equation can now be formulated in terms of moments of the soot particle size distribution using the same procedure as for the particle inception:

$$\dot{M}_{r,com} = A - B$$

$$A = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left((i+j)^r \beta_{i,j} N^p{}_j N^s{}_i\right)$$

$$B = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(i^r \beta_{i,j} N^s{}_i N^p{}_j) \ . \quad (11)$$

Under the assumption that soot particle are much larger than PAH-molecules (i>>j) the collision frequency $\beta_{i,j}$ can be approximated eq.(6) as:

$$\beta_{i,j} = C_{nn} j^{\frac{1}{2}} i^{\frac{2}{3}}, \ (i>>j). \quad (12)$$

So that the source terms of condensation for the *r-th* moment becomes:

$$\dot{M}_{r,con} = C_{nn}\sum_{k=0}^{r-1}\binom{r}{k}\left(M^s{}_{k+\frac{2}{3}} M^p{}_{r-k-\frac{1}{2}}\right). \quad (13)$$

D. Surface Growth and Oxidation

The rate surface growth was empirically determined by Wagner [8], who showed that it can be expressed by a first-order rate law in soot volume fraction, k being an empirical rate constant:

$$\frac{df_v}{dt} = k(f_{v,\infty} - f_v) \quad (14)$$

A slight different growth mechanism as applied in the soot model used here, see Table 1. The main difference to the HACA-mechanism is that the ring-closure step is divided into two individual reactions. First acetylene is added (3.a) and in the second step (3.b) the ring is closed. The addition of acetylene is assumed to be reversible while the ring closure is irreversible. This is due to the finding that the bound between the acetylene and the soot surface can be broken at high temperature. This shall explain the low level of soot formation at location in the flame where temperatures are high.

TABLE 1 THE DEVELOPED HACA-MECHANISM

(1.a)	$C_{soot,i}H + H \xleftrightarrow{k_{1A}} C^*_{soot,i} + H_2$
(1.b)	$C_{soot,i}H + OH \xleftrightarrow{k_{1b}} C^*_{soot,i} + H_2O$
(2)	$C^i_{soot-*} + H \xleftrightarrow{k_2} C^i_{soot-H}$
(3.a)	$C^i_{soot,i} + C_2H_2 \xleftrightarrow{k_{3a}} C^*_{soot,i}C_2H_2$
(3.b)	$C^i_{soot,i}C_2H_2 \xleftrightarrow{k_{3b}} C_{soot,i+1}H + H$
(4.a)	$C^*_{soot,i} + O_2 \xrightarrow{k_{4a}} C^*_{soot,i-1} + 2CO$
(4.b)	$C^*_{soot,i}C_2H_2 + O_2 \xrightarrow{k_{4b}} C^*_{soot,i} + 2CHO$
(5)	$C_{soot,i}H + OH \xrightarrow{k_5} C^*_{soot,i-1} + CH + CHO$

The reaction rates are assumed to be those of the analogous gas phase reaction of phenyl and benzene. The reaction rate of reaction (3.a) for particles of size class *i* can thus be expressed as:

$$r_{3a,i} = k_{3a,f}[C_2H_2][C_{soot}] \tag{15}$$

A steady-state assumption for the active radical sites [C^*_{soot}] and [$C^*_{soot}C_2H_2$] is introduced, leading to algebraic equation for the concentration of active radical sites:

$$[C^*_{soot}] = [C_{soot}] \cdot \frac{A}{B} \tag{16}$$

$$[C^*_{soot}C_2H_2] = \left(\frac{k_{3a,f}[C_2H_2]}{k_{3b,f} + k_{3a,b} + k_{4b}[O_2]} \right) \cdot [C^*_{soot}] \tag{17}$$

with

$$A = k_{1a,f}[H] + k_{1b,f}[H] + k_5[OH] \tag{18}$$

$$B = k_{1a,b}[H_2] + k_{1b,b}[H_2O] + k_2[H] + k_{3a,f}[C_2H_2]f_{3a} + k_4[O_2] \tag{19}$$

$$f_{3a} = \frac{k_{3b,f}}{k_{3b,f} + k_{3a,b} + k_{4b}[O_2]}. \tag{20}$$

If the ring closure reaction (3b,f) is fast compared to the fragmentation (3a,b) and the oxidation (4b) the variable f_{3a} is close to 1. In this limit the expression above is identical to that obtained from the HACA-mechanism. At the other extreme, if fragmentation becomes the dominant process f_{3a} gets small and the growth process is stopped. The concentration of active sites can be calculated by:

$$[C_{soot}] = \sum_{i=1}^{\infty} \alpha \frac{\chi_{soot}}{N_A} S_i N_i \ . \tag{21}$$

Where α is the fraction of surface sites, and χ_{soot} is the number of surface sites per unit area, and S_i is the surface area and N_i is the number density of particles of size class *i*. Since benzene structure on the soot surface has one active site ($\chi_{soot}\, S_1$), the following expression is valid:

$$\chi_{soot} S_i = \chi_{soot} S_1 i^{2/3} = i^{2/3} \ . \tag{22}$$

Now the source terms of surface growth and oxidation for the moments of soot can be formulated in general for surface reactions:

$$\dot{M}_{0,sg} = 0 \tag{23}$$

$$\dot{M}_{r,sg} = \alpha\, k_{3a,f}[C_2H_2] f_{3a} \frac{A}{B} \sum_{k=0}^{r-1} \binom{r}{k} M^{\,s}{}_{k+\frac{2}{3}} 2^{r-k}, \quad r = 1,2,\ldots \tag{24}$$

$$\dot{M}_{0,ox} = -\alpha \left(k_{4a}[O_2]\frac{A}{B} + k_5[OH] \right) N_x \tag{25}$$

$$\dot{M}_{r,sg} = \alpha \left(k_{4a}[O_2]\frac{A}{B} + k_5[OH] \right) \sum_{k=0}^{r-1} \binom{r}{k} M^{\,s}{}_{k+\frac{2}{3}} 2^{r-k} \quad r = 1,2,\ldots \tag{26}$$

The system of equations is not closed since the number density of the smallest size class N_1 is not known. It is therefore assumed that the probability of the burn-out of the soot particles is proportional to the mean particle size. The rate of oxidation for the zero-th moment writes after weighing with the mean number of C-atoms as:

$$\dot{M}_{0,ox} = -\left(k_{4a}[O_2]A + k_5[OH] \right) M^{\,s}{}_{-\frac{1}{3}} \ . \tag{27}$$

E. Soot Source Terms

The formation of soot can be subdivided into the process of particle inception, condensation, surface growth oxidation. As for the PAHs a statistical approach is used to describe the size distribution function of the soot particles. The moments of the size distribution of the size are defined as:

$$M_r = \sum_{i=1}^{\infty} i^r N_i \qquad r = 0,1,\ldots, \tag{28}$$

Where Ni is the number density of particle i with a mass $m_i = i \cdot m_1$, with m1 being the mass of the smallest unit occurring in a soot particle. The moment M0 is equal to the total particle number density:

$$M_0 = \sum_{i=1}^{\infty} N_i = N \ . \tag{29}$$

The moment M1 can be related to the volume fraction, which defines the ratio of the volume occupied by soot particles to the gas volume:

$$M_1 = \sum_{i=1}^{\infty} i N_i = f_v \frac{\rho_s}{m_1} \ . \tag{30}$$

The source term $\dot{M}_r$ is:

$$\dot{M}_r = \dot{M}_{r,pb} + \dot{M}_{r,con} + \dot{M}_{r,sg} + \dot{M}_{r,ox} \tag{31}$$

with pb= particle inception, con=condensation, sg=surface growth, ox=oxidation.

F. Application to Turbulent Combustion

The soot moment transport equation [6] is employed in this study:

$$\rho \frac{\partial M_r / \rho}{\tau} - \rho \frac{\chi}{2} \frac{\partial^2 M_r / \rho}{\partial Z^2} - \dot{M}_r = 0 \qquad i = 1,\ldots,\infty. \tag{32}$$

Where Z is mixture fracture, and χ is scale dissipation rate, and τ is time. Some data on the engine is summarized. The test engine is the four cylinder, 2.8 L 4JB1 DI engine. The injection system is a third generation Bosch Common-Rail featuring a maximum injection pressure of 1600 bar.

Figure. 1 displays the source terms (oxidation not included) to the first soot moment as a function of time resulting from the simulation of the combustion process in a DI diesel engine. Injection startted at 2.0 degrees before *TDC* and the first soot starts to form in small amounts at approximately 2.0 degrees after *TDC*. (The source terms due to particle coagulation has been omitted, since this process does not contribute to the growth in soot volume fraction, and is thus equal to zero at all times). It is seen that, apart from the initial phase of particle inception, condensation of PAH onto the soot particle surface is the dominating process. It should be emphasized that this process is treated seperately from surface growth in the current model.

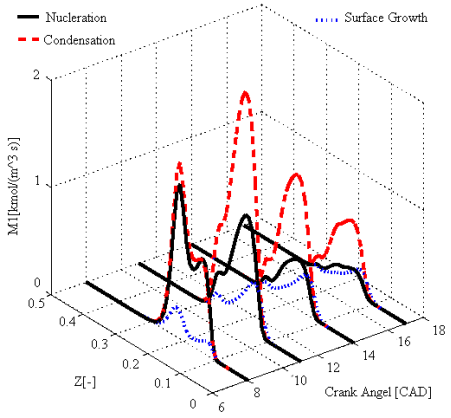


Figure. 1. Soot source terms in a DI diesel engine

Figure. 2 shows the total soot as a function of engine crank angle and the source terms owing to oxidation by OH-radicals and molecular oxygen,

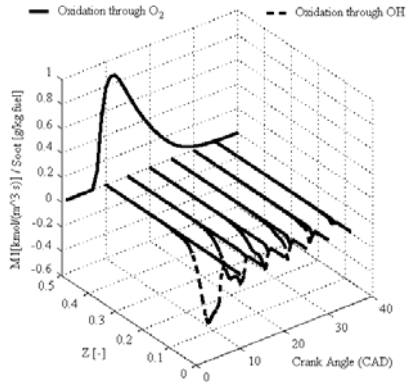


Figure. 2. Soot oxidation source terms and total soot

respectively. The figure clearly demonstrates that oxidation by OH-radicals is the dominating process under conditions relevant for diesel engine combustion. Total soot as a function of engine crank angle is also shown. The final soot volume fraction generally depends on pressure, temperature, mixture composition, and the fuel structure. Engine-out soot levels are obtained as the difference between two large numbers, representing the processes of formation and oxidation, respectively. This is one of the inherent problems making quantitative soot predictions so challenging. The simulations suggest that, in this case, only around 50% of the soot formed gets subsequently oxidized. This should be compared to almost 90-96% for the most advanced injection timings. The results are similar to that observed by Schwarz et al. [11], who performed optical diagnostics in a transparent heavy-duty DI diesel engine. The fact that less soot is formed is in this case a result of incomplete combustion, which is confirmed by the experiments, shows a significant increase in unburned hydrocarbons and carbon monoxide.

III. Conclusions

(1)Apart from the initial phase of particle inception, condensation of PAH onto the soot particle surface is the dominating process.

(2) Oxidation by OH-radicals is the dominating process under conditions relevant for diesel engine combustion.

Acknowledgment

This project has been supported by Beijing Municipal Commission of Education (KM200910016014) and Ministry of Housing and Urban-Rural Development of the People's Republic of China (MOHURD) (2009-K8-5).

References

[1] Liu，Yongfeng ,Pei，Pucheng, Asymptotic analysis on autoignition and explosion limits of hydrogen-oxygen mixtures in homogeneous systems, International Journal of Hydrogen Energy[J], 2006, 31(5),639～647

[2] Peters, N., Turbulent Combustion [M], Cambridge University Press, Cambridge, 2004, 1-5

[3] Balthasar, M. et al, Implementation and validation of a new soot model and application to aero engine combustors, ASME [J], 2000-GT-0142

[4]Mauss,F. Entwicklung eines kinetischen Modells der Russbildung mit schneller Polymerisation, Ph.D. Thesis[M], RWTH Aachen, 1997

[5] Balthasar，M., Heyl A., Mauss F., Flamelet modeling of soot formation in larminar ethyne/air diffusion flames, Twenty-sixth symposium (international) on combustion, Naples[C], 1996, 2396-2405

[6] Frenklach,M. and Wang,H. Soot formation in combustion-mechanisms and models[M], Springer Verlag, Berlin-eidelberg, 1994, 162-164

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 062-065

Strong Barrier Coverage for Intrusion Detection in Wireless Sensor Network

Jianbo Li[1], Shang Jiang[2], and Zhenkuan Pan[3]
Information Engineering College, Qingdao University, Qingdao, 266071, Shandong Province ,P.R.China
Email: [1]lijianboqdu@yahoo.com.cn, [2]jiangshan66@163.com, [3]zkpan@qdu.edu.cn

***Abstract*—Due to the feature of self-organization and self-configuration of wireless sensor networks, it is very suitable for the application of intrusion detection, such as national boundary monitoring. As for the cost of sensor nodes, deriving the critical density is a fundamental problem for intrusion detection. Recently there has been a lot of research on the problem of intrusion detection by modeling it as a barrier coverage problem. Different from previous works on the distribution of sensor nodes, which usually lead to an upper bound of density, we try to get the critical density for intrusion detection. In this paper, we model the problem of intrusion detection as a strong barrier coverage problem and solve it by a percolation method. We prove that the critical node density for intrusion detection is 2ln2. When the node density is smaller than 2ln2, there is always a crossing path that cannot be detected by sensor networks. By contrary, when the density is larger than 2ln2, there are, with high probability, no such paths.**

***Index Terms*—critical node density, intrusion detection, strong barrier coverage, wireless sensor network**

I. INTRODUCTION

A wireless sensor network consists of a large number of sensor nodes equipped with RF radios. Sensor nodes can communicate with the far away base station through multi-hop transmissions. Due to their intrinsic self-configuration and self-organization feature, wireless sensor networks are very suitable for the application of intrusion detection, such as national boundary monitoring. In a national boundary monitoring application scenario, sensor nodes are usually randomly deployed on the boundary of a country. Intrusion warnings are reported to the base station once sensors have detected intruders.

Recently, there have been a lot of works on the problem of intrusion detection. Most of them model the intrusion detection problem as a barrier coverage problem. Many algorithms have been proposed for the intrusion detection problem; however, they mainly focus on the distribution of sensor nodes, which usually lead to an upper bound of the density of sensor nodes. In a practical deployment, the deployment of sensor network is usually required to guarantee that all intrusion warnings should be reported to the base station. That is, there should be no crossing paths that are not monitored by any sensor node. Taking the cost of deploying sensor nodes into consideration, the critical node density is a desired value for a practical deployment.

In this paper, we model the intrusion detection problem as a strong barrier coverage problem and try to derive the critical density of sensor nodes by using a percolation method. This critical density guarantees that when the density of deployed sensor nodes is larger than the critical node density, all intrusions should be detected and reported to the base station. The critical density of sensor nodes is a result in the meaning of statistics and has twofold meanings. On the one hand, when the density of sensor nodes is higher than the critical density, it guarantees that there are, with high probability, no crossing paths which are not detected by any sensors. On the other hand, when the density of sensor nodes is smaller than the critical density, there is always a crossing path with high probability.

Our contributions in this work are mainly as follows:

1) We model the intrusion detection problem as a strong barrier coverage problem and solve it by a percolation method.
2) We derive the critical density of sensor nodes. We also validate our result by simulations. The simulation results show that our theoretical critical density value is very close to the practical critical value.

The rest of this paper is organized as follows: in Section II, we present the works related to our work. We introduce the network model and assumptions used in this paper in Section III. Section IV presents our critical density value and its proof. We validate our theoretical critical density value by simulations in Section V. Finally, we conclude this paper in Section VI.

II. RELATED WORKS

In this part, we review the previous works on barrier coverage. The problem of barrier coverage is first proposed by Gage [1]. The purpose of barrier coverage is to detect intruders who attempt to cross from one side to another side of the monitoring area. Several different measurements have been proposed for barrier coverage.

The measure of path coverage is proposed in [2] and efficient algorithms are presented to find the maximum breach paths and maximum support paths. In [3], the measure of path exposure is proposed to measure the probability of detecting an intruder when the intruder moves along a given path. A centralized algorithm is proposed to find the minimum exposure paths, which have the minimum probability of being detected. The distributed algorithms are proposed for these path coverage problems in [4, 5].

Liu and Towsley [6] studied the barrier coverage problem in two-dimensional plane and two-dimensional strip. They investigate three coverage measures (area

AP-PROC-CS-09CN005

coverage, node coverage and detectability) using percolation theory results.

Kumar et al. introduce strong barrier coverage and weak barrier coverage in [7]. In strong barrier coverage, all intrusions will be detected. That is, no matter which the intrusion path is selected, intruders will be detected. However, in weak barrier coverage, there exist some crossing paths along which intruders can cross the monitoring area without being detected. Fig.1 gives an example of strong barrier coverage. Fig.2 gives an example of weak barrier coverage. A centralized algorithm is proposed in [7] to determine whether a region is *k*-barrier covered that all intrusions are sensed by at least *k*-different sensor nodes. They also derive the critical conditions for weak barrier coverage in a randomly deployed sensor network.

As the centralized algorithm may incur heavy communication loads and computation costs on the sensor nodes, Chen et al. [8] design a localized algorithm to detect intrusions in a belt region of deployment. The belt region is partitioned into several pieces of length *L*. Sensor nodes in the same piece cooperatively check that whether the local piece is *k*-barrier coverage. After that, they exchange the check results. When all pieces of the belt region is *k*-barrier coverage, the belt region is *k*-barrier coverage; Otherwise, it is not.

Ai Chen et al. [10] study the quality of barrier coverage. Different from works that evaluate the barrier coverage quality of a sensor network by the criterion that whether it can guarantee that this is no crossing path, they propose the measurement of *L*-local *k*-barrier covered to qualify the identified area, which is much fairer to evaluate the barrier ability of sensor networks.

As all these works mentioned above are mainly focusing on the distribution of sensor nodes after deployment, they usually lead to the upper bound of node density. In this paper, we focus on the strong barrier coverage problem and try to derive the critical density for intrusion detection.

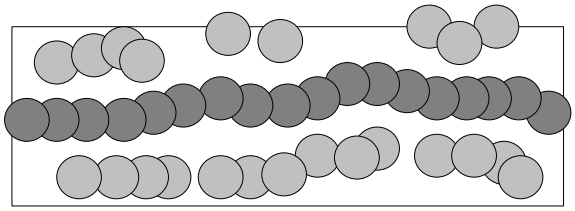


Figure1. Strong barrier coverage: all black nodes form a sensor barrier that guarantees that no intruders can cross the monitoring area from upside to downside.

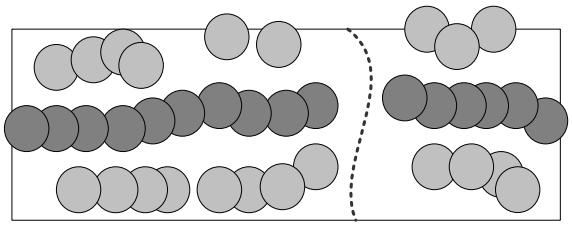


Figure2. Weak barrier coverage: a crossing path helps intruders to cross the monitoring area without being detected.

III. Network Model

In this part, we present our model and assumptions for the intrusion detection problem.

We consider the intrusion detection problem in a thin strip rectangle area of size $A=l\times w$, where l is the length of the strip area and w is the width of the strip area. Sensor nodes are randomly deployed in this area according to Poisson point process of density λ. That is, the expected number of sensors in this area is λlw. We assume that all sensors are static after deployment.

We use the widely adopted Boolean sensing model. Under this model, each node has a certain sensing range r. A sensor can only sense the environment and detect intruders within the sensing range r. A location is said to be covered only when it is within the sensing range of some sensor node. After the deployment of sensor nodes, the monitoring area is divided into two parts, one is the area covered by sensor nodes, and the other is uncovered area.

As warnings are usually required to be reported to the base station, we assume the communication range of each node, r_c , is at least two times of the sensing range r. That is, $r_c \geq 2r$. We say two sensor nodes are overlapped or connected when their Euclidean distance is smaller than $2r$. For simplicity, in this paper, we normalize the sensing range r as 1.

A sensor barrier is defined as a connected component of sensors which connects both the left side and right side of the monitoring area as illustrated in Fig. 1. It can be found that an intruder cannot cross the monitoring area from upside to the downside without being detected by the sensor barrier.

A crossing path is a path connecting the upside and the downside of the monitoring area, where the ingress point and the egress point are on the upside and downside of the area, respectively. When an intruder moves along the crossing path, he cannot be detected by the sensor network. It can be implied that every point in the path is not covered by any sensor node. Fig.2 gives an example of a crossing path.

The detectability of a sensor network can be measured by the criterion that whether it guarantees that all intrusions are detected. That is, if there is a sensor barrier, the detectability of that sensor network is good. Otherwise, it is bad.

Definition: A sensor network is said to be strong barrier covered only when P(any crossing path is at least 1-barrier covered) = 1.

IV. Critical Density For Intrusion Detection

In this section, we present the critical density for intrusion detection problem.

Theorem 1: Consider a sensor network randomly deployed in a strip area of size $A=l\times w$ according to the Poisson point process of density λ, where l and w is the

length and width of the strip area, respectively, the critical density of sensor nodes is about 2ln2.

Proof: We first convert the intrusion detection problem to a bond percolation problem and use the result presented in [11].

We divide the monitoring area into small equal squares with length of $1/\sqrt{2}$, as illustrated in Fig. 3.

As the density of Poisson process is λ, the probability that there is no sensor in a small square is $p = e^{-\lambda/2}$. A square is said to be open if it is not covered by any sensor node, otherwise, it is said to be closed. When a square is open, there exists a path crossing that square. This construction can be mapped to a bond percolation model as follows. Horizontal edges and vertical edges are added between adjacent small squares, as shown in Fig. 4.

A path consists of a sequence of consecutive edges and is said to be closed when one of its edge is closed. A crossing path is defined as a path consisting of only open edges.

As for that the critical value of a bond percolation problem is 1/2 as presented in [11], $p = 1/2$. So $\lambda = 2ln2$. ■

By percolation theorem, when the open probability is larger than the critical value, with high probability, there is a path from one side to the other side. When the open probability is smaller than the critical value, with high probability, there is no such path. So when the node density is larger than the critical density, there are crossing paths from the upside to the downside while there is no crossing path when the node density is smaller than the critical density.

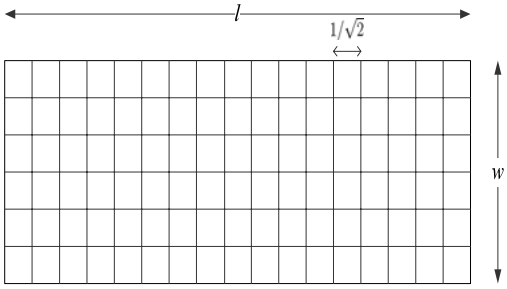


Figure3. The area is partitioned into equal small squares of length $1/\sqrt{2}$

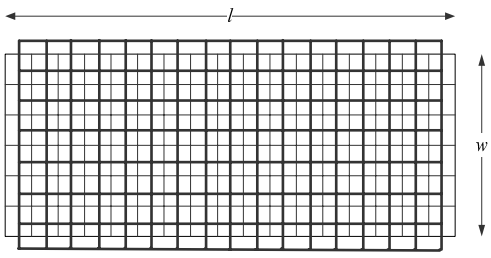


Figure4. We convert intrusion detection problem to a bond percolation problem. Edges are added between adjacent small squares.

V. Performance Evaluation

In this section, we validate our critical node density result by simulations. We randomly deploy sensors in a strip area. The sensing range of each sensor node is set as $1m$ and the communication range is set as $2m$. We test different node density and check whether the sensor network is at least 1-barrier coverage. To get a fair evaluation, we generate 20 different networks for the same density parameter.

A. The Impact of Node Density

In this part, we test the impact of node density on the intrusion detection. We deploy sensor nodes randomly in a strip area of size $10m \times 100m$. Fig.5 shows the results with the number of sensor nodes varying from 100 to 1500. The blue line in Fig. 5 shows the percent of 1-barrier coverage in 20 different topologies while the red line is the theoretical critical value. We can find that when the number of node is larger than about 900, all topologies are at least 1-barrier coverage while none of topologies are 1-barrier coverage when the number of node is small than about 500. This phenomenon validates our result that there exists a critical value that when the node density is larger than the critical value, with high probability, all networks are at least 1-barrier coverage, while no 1-barrier coverage exists when the node density is smaller than the critical value.

Note that there is a gap between them. When the number of sensor nodes is larger than about 900, roughly all topologies have at least one 1-barrier coverage that guarantees that all intrusions will be reported, however, the critical value is about 1386. It is because that we partition the monitoring area by small squares while the sensing range is a disk by the sensing model. As our partition is just a conservative approximation for the sensing range, our theoretical critical value is a little greater than the practical value.

B. The Impact of Strip Length

In this part, we investigate the impact of the length of the strip area on the critical number of nodes with length varying from 100m to 1000m in Fig.6. The width of the

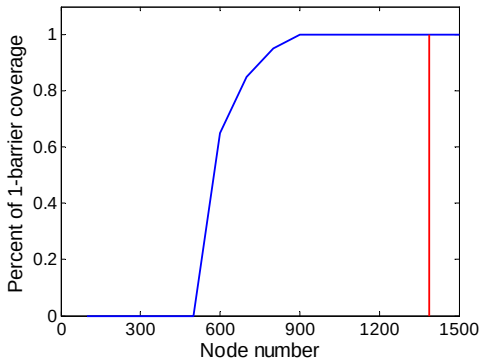


Figure5. The percent of 1-barrier coverage when the number of node is varied from 100 to 1500. The red line is its theoretical value. The blue line shows the percent of 1-barrier coverage in 20 different topologies with different number of node.

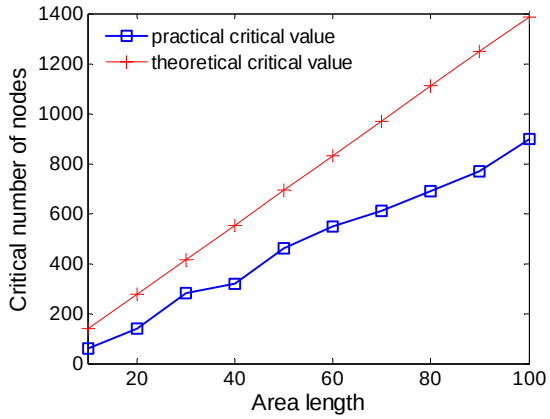


Figure6. Comparison between practical critical value and theoretical critical value with the length of area varying from 10m to 100m

strip area is fixed as 10m. We compare the practical critical values with our theoretical values in different lengths of the strip area. As there is a phase from that there is no 1-barrier coverage to that there is a 1-barrier coverage as illustrated in Fig. 5. We only consider the minimum number of nodes with the full percent of 1-barrier coverage in 20 different topologies.

Fig.6 gives the critical number of nodes in different length of the strip area. We can find that the practical critical value is always smaller than the corresponding theoretical value with the same reason as explained above. Note that the practical critical value is almost increased linearly as shown in Fig. 6.

VI. Conclusion

In this paper, we study the problem of intrusion detection and model it as a bond percolation problem. Different from previous works which are usually conservative and lead to an upper bound of node density for guaranteeing that there are no missing alarms, we obtain the critical density value for intrusion detection. We validate our result by simulations, and the performance result shows that when the node density is larger than the critical density, (even it is a litter higher than the critical density), there are no crossing path in the deployed sensor network.

References

[1] Gage. Command control for many-robot systems. In Proc. of the Nineteenth Annual AUVS Technical Symposium (AUVS-92), 1992.

[2] S. Meguerdichian, F. Koushanfar, M. Potkonjak, and M. B. Srivastava. Coverage problems in wireless ad-hoc sensor networks. In Proc. IEEE Infocom, pages 1380-1387, 2001.

[3] S. Meguerdichian, F. Koushanfar, G. Qu, and M. Potkonjak. Exposure in wireless ad-hoc sensor networks. In ACM Mobile Computing and Networking, pages 139-150, 2001.

[4] X.-Y. Li, P.-J. Wan, and O. Frieder. Coverage in wireless ad-hoc sensor networks. IEEE Transactions on Computers, 52(6):753-763, June 2003.

[5] G. Veltri, Q. Huang, G. Qu, and M. Potkonjak. Minimal and maximal exposure path algorithms for wireless embedded sensor networks. In Proc. of ACM Sensys, 2003.

[6] Liu and D. Towsley. A study on the coverage of large-scale sensor networks. In The 1st IEEE International Conference on Mobile Ad-hoc and Sensor Systems, 2004.

[7] S. Kumar, T. H. Lai, and A. Arora. Barrier coverage with wireless sensors. In Proc. ACM Mobicom, 2005.

[8] Chen, S. Kumar, and T.-H. Lai. Designing localized algorithms for barrier coverage. In Proceedings of ACM Mobicom, 2007.

[9] Benyuan Liu and, Olivier Dousse, Jie Wang and Anwar Saipulla, Strong barrier coverage of wireless sensor networks. In Proceedings of ACM MobiHoc, 2008.

[10] Ai Chen, Ten H. Lai and Dong Xuan, Measuring and Guaranteeing Quality of Barrier-Coverage in Wireless Sensor Networks. In Proceedings of ACM MobiHoc, 2008.

[11] Geoffrey Grimmet, Percolation, Grundlehren der mathematischen Wissenschaft, vol 321, Springer, 1999.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 066-070

A PSO Algorithm Based on Biologe Population Multiplication (PMPSO)

Lei Yin[1], and Xiaoxiang Liu[2]
[1] School of Mechano-Electronic Engineering Xidian University, Xi'an, China
Email: yinlei_w@163.com
[2] Department of Computer Science of Zhuhai College Jinan University, Zhuhai, China
Email: tlxx@jnu.edu.cn

***Abstract*—Inspired by the natural phenomenon of multiplication of biological population, a population multiplication particle swarm optimization (PMPSO) is presented. The proposed algorithm (PMPSO) has four phases of migration, selection, elimination and reproduction, evolution. Using searching optimal model of PSO in the migration phase; introducing LEVEL SET theory dividing population to be able to facilitate the selection operation in the selection phase; speeding up the algorithm convergence by abandoning the inferior population, reproducing superior population and making full use of population resource in the phase of elimination and reproduction; creating new population to keep the diversity to avoid monotone of the algorithm in the last evolutionary phase. Finally, PMPSO is applied to some test functions comparing with GA and SPSO algorithm, which is proved that the PMPSO is feasible and effective.**

***Index Terms*—biologe population multiplication, LEVEL SET, Particle Swarm Optimization**

I. INTRODUCTION

The particle swarm optimization (PSO) algorithm originally was developed by Kennedy and Eberhart in 1995 [1]. PSO is suitable to both scientific research and engineering applications [2]. Moreover very few parameters are needed to be adjusted, which makes it particularly easy to implement. However, it is pointed out that although PSO can show significant performance in the initial iteration, it might encounter problems in reaching optimum solutions efficiently for several approximation problems. It is obvious that the particle swarm loses its diversity and all the particles are attracted towards the best position so far by any of particles.

A lot of research work is made in order to overcome the disadvantage of PSO. Ref. [3] proposed a 'stretching' function, which consists of a two-stage transformation of the objective function, to alleviate the local minima problem. Ref. [4] presented a predator prey model to maintain diversity in the swarm and prevent premature convergence to local minimum. Ref. [5] introduced a PSO model with passive congregation to help individuals to avoid misjudging information and becoming trapped by poor local minima. Other studies on dealing with this issue were undertaken using multiple populations in [6] and survival density concept in [7].

corresponding author: tlxx@jnu.edu.cn

AP-PROC-CS-09CN005

In nature, each population will search food in order to multiply. As we all know that the rule of survival of the fittest, original but effective, exits in the process of searching food. In this paper, we introduced this rule to PSO algorithm eliminating inferior population and keeping superior population. It is helpful to make full use of population resources and speed up the algorithm convergence. In the selection phase, classifying successfully by using LEVEL SET theory make the algorithm accord with the principle of survival of the fittest. At the same time, we also take into account the evolution of population to keep the diversity of the population which can prevent the monotone and prematurity of the algorithm. Finally, the algorithm is applied to some test functions to verify its feasibility and effectiveness.

II. A PSO ALGORITHM

A. Standard particle swarm optimization (SPSO)

PSO was presented by Kennedy and Eberhart [1] in 1995. In the PSO system, a number of particles coexist and cooperate to find optimization. Each particle "flies" to a better position in problem space in accordance with its own "experience" and the best "experience" of the adjacent particle swarm, searching the optimal solution.

Mathematical notation of PSO is defined as follow:

Assume searching space is D-dimensional and the total number of particles is n. The ith particle location is denoted by the vector: $X_i=(x_{i1}, x_{i2},\ldots, x_{iD})$; The past optimal location of the ith particle in the "flight" history (that is, the location corresponds optimal solution) is $P_i=(p_{i1}, p_{i2},\ldots, p_{iD})$. The past optimal location P_g of the gth particle is optimal in all of $P_i(i=1,2,\ldots,n)$; The location changing rate (speed) of the ith particle is denoted by the vector $V_i=(v_{i1}, v_{i2},\ldots, v_{iD})$.The location of each particle changes by the following formula:

$$v_{id}(t+1)=wx_{id}(t)+c_1 rand()(p_{id}(t)-x_{id}(t)) + c_2 rand()(p_{gd}(t)-x_{id}(t)) \quad (1)$$

$$X_{id}(t+1)=x_{id}(t)+v_{id}(t+1),\ (1\le i\le n,\ 1\le d\le D) \quad (2)$$

c_1, c_2 are positive constants called accelerating factor; $rand()$ is a random number between 0 and 1; $[X_{\min}, X_{\max}]$ is the changing range of particle location. $[v_{\min}, v_{\max}]$ is the changing range of speed. If the location and speed exceed boundary range in iteration, given boundary value. w is called inertia factor; w, setted a litter bigger ,is suited

to a wide range of exploration to solution space while smaller is suited to a small range.

B. The shortcomings of conventional PSO algorithm

As shown in Fig. 1, each particle of PSO closes to historical optimal location and global optimal location. This makes PSO algorithms have many advantages, such as that their computational complexity doesn't increase with the rising of the dimension of the problem, and rapid convergent speed, etc. However, they still have some shortcomings, which are listed as follows:

Shortcoming 1: When the conventional PSO searches, the particles tend to get close to the better particles. This property would make the algorithm find out the optimal solution as soon as possible, however, this property is also a flaw that could result in premature convergence. That is, when all the particles constantly get close to the better ones, all the particles in the system would be probably concentrated in a local optimal solution. At this situation, it is a pity that all the particles can not jump out of the local optimal solution they have approached. Fig. 2 illustrates such phenomenon:

From Fig. 2, it can be seen clearly that particles don't find the global optimal solution but concentrate to a local optimal solution. At this time, they no longer have the abilities to get rid of the attraction of the local optimal solution, and result in premature convergence.

Shortcoming 2: The speeds of particles are too great. When particles are located in some local, the objective function is quite sensitive to the slight changes of particles. Thus, at this time, too great speed of the particle is not suitable; meanwhile, too little speed would influence the speed of convergence.

We can see from Fig. 3 that though particle is attracted by the optimal solution, and motion toward the optimal solution. Nevertheless, because the speed of particle is too great, it would easily miss the optimal solution.

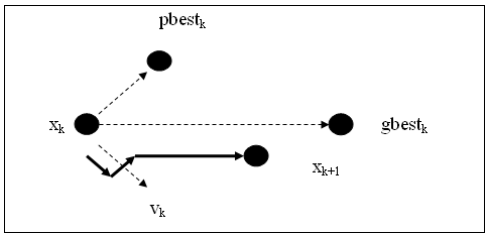


Figure 1. sport of the particle

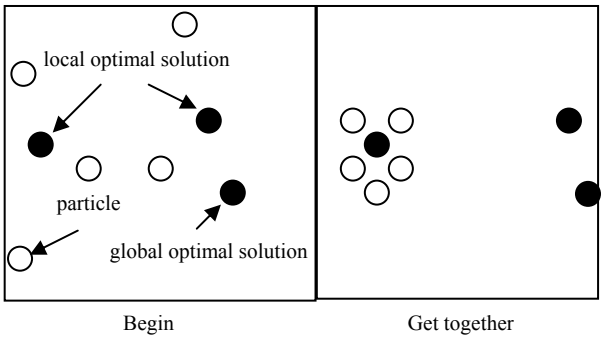


Figure 2. particle get together

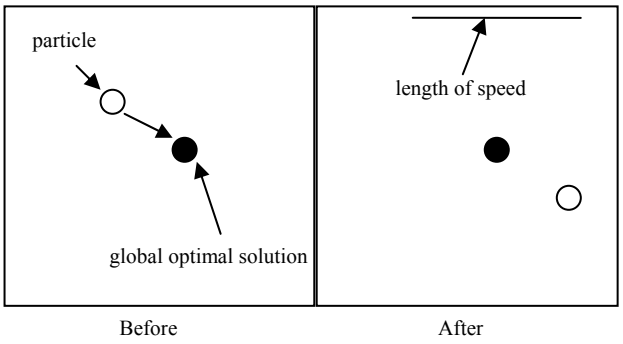


Figure 3. particle moves to the best position

The above shortcomings in PSO algorithm are like some flies in the ointment. To make up those weaknesses existing in PSO, the following text would give some concrete schemes, which include MDPSO algorithm to overcome the first shortcoming. as well as numerical-level weight to control the speed for deal with the second shortcoming.

III. A PSO ALGORITHM BASED ON BIOLOGICAL POPULATION MULTIPLICATION

A. Biological population multiplication

In nature, populations search food in order to multiply. As we all know the rule of survival of the fittest, original but effective, exits in the process of searching food.

First of all, we assume that some biomes are dotted in a region. Each of them migrates to search food as well as a more suitable place for survival. In the Fig. 4, this article assumes that there are four communities, *p*1, *p*2, *p*3, *p*4, in a region, Because of the need looking for food, community migration is called respectively: *P*1, *P*2, *P*3, and *P*4. And after that, the survival of the fittest begins. Among them, *P*3 and *P*4 successfully accepted the test to continue to survive, besides *P*3 takes further reproduction to extend the community due to good environment; *P*1, tortured by the nature, evolves eventually to become *P*1' adapting to the environment; but *P*2 has to be eliminated because it is hard to find suitable places to survive. This mode of biomes multiplication not only washes out the inferior population and keeps the superior ones, but also stimulates the evolution of population to adapt to the survival environment. For this right mode, hundreds of thousands of biological communities could survive and continue.

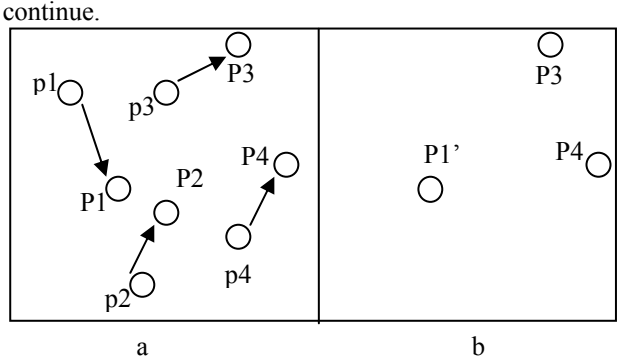


Figure 4. particle survival

B. *Improved PSO Algorithm Based on the Population Multiplication*

We know that in PSO algorithm, each particle moves towards the global optimal location and the optimal location of individual history as a criteria to find a better location for survival. This model allows algorithm has a good convergence, but also maintains a good searching performance. In Fig. 5, after a round of movement, the particles all have new locations A1, B1, C1, D1, E1 and F1. However, each new particle continues to search optimization directly without the process of survival of the fittest in the next movements, illustrated as shown in Fig. 6.

However, this movement in PSO makes some inferior particles continue to reproduce to become inferior communities unable to be eliminated which affects the algorithm convergence rate. At the same time the resource of particle swam can not be fully utilized. That is because the quantity of particles affects the algorithm efficiency while the quality of particles does the same. In order to overcome this disadvantage, the paper presented an improved PSO algorithm with the principles of biologic population multiplication. The algorithm is divided into four phases: migration, selection, elimination and reproduction, evolution.

1) Migration

We introduce the concept of migration to the new algorithm. The population migration is similar to the changes of the particles location in PSO, one changes for the survival of population while the other is for a better location. And the migration of population is also affected by two factors: history experience and communication experience. The history experience just means searching the optimal location of individual and communication experience is for the global optimal location in PSO. So

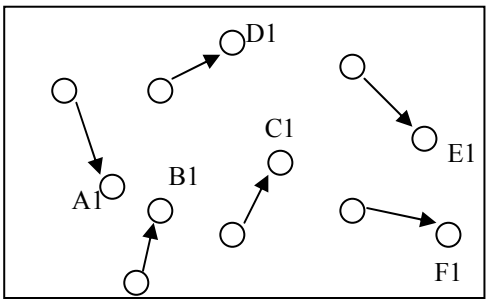


Figure 5. The movement of first generation particles, each of them moves to search a better place

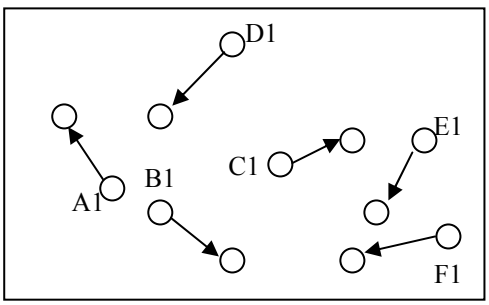


Figure 6. Traditional PSO algorithm: Each particle gets location of the next particle after the previous round and continues to move

at this phase, the new algorithm and PSO algorithm look like the same (maybe only the name is different). We will still use the speed changing (1) and location changing (2) of PSO.

In the (1), the value w is fixed. w, set a litter larger, is suit to a wide range of exploration to solution space while smaller is suit to a small range. At the early convergence, the larger w can speed up the convergence, while in the latter the smaller w can improve the capacity of searching optimization. Therefore, this paper defines the w as follow:

$$w(i)=w_{\max}-i(w_{\max}-w_{\min})/N$$

Here, $w(i)$ is alterable (maybe degressive more exactly), $w_{\max}$, $w_{\min} \in (0, 1)$.

2) Selection

At selection phase, we need to judge which population will be eliminated and reproduce and how much they reproduce. This requires that all population should be divided into two parts: the superior ones and the inferior ones. LEVEL SET theory is introduced here.

For the tth-generation $P(t) = (P1, P2, ..., Pn)$, n denots the number of particles, the fitness function of of particles is set to $f_i(x)$, order

$$\overline{f_t} = \sum_{i=1}^{n} \frac{f(X_i)}{n}$$

$$H_{\overline{f_t}} = \{x_i \in P(t) \mid f(x_i) \le \overline{f_t}, 1, 2, \cdots, n\}$$

Where t denotes tth-generation. $H_{\overline{f_t}}$ is called the level set about f relative to $P(t)$. After that the population of each generation can be divided two parts [8].

Selection steps are as follows:

a) Set the initial population for $X = (X_1, X_2, ..., X_n)$;

b) Calculate the fitness of each population;

c) Calculate the mean of fitness $\overline{f_t}$, in which the population better than $\overline{f_t}$ belongs to X_a, and the worse ones belong to X_b.

d) According to the method of Step *c)*, X_b is divided into X_c and X_d, between them X_c stands for the better population, and X_d for the poorer population.

e) The population number in X_d is *nd.* So we select randomly *nd-pm* in $X_a+X_b+X_c$ for reproducing. *pm* is the number of evolution population discussed below.

3) Elimination and Reproduction

When population arrives in a new environment, which is too bad to adapt to, the entire population has to be extinct which is called elimination. However, when they arrive an eminent environment, they will be developed and reproduce. This concept introduced in new algorithm is completely different with the PSO algorithm. Fig. 6 has illustrated the particle change of PSO algorithm, changes of the improved PSO algorithm is as follows in Fig. 7:

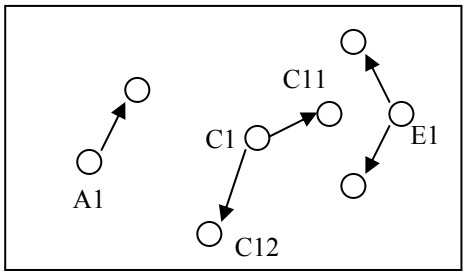


Figure 7. particle movement in the PMPSO

The difference between Fig. 6 and Fig. 7 is B1, D1, and F1 are all eliminated and disappear while A1, C1, and E1 take a further reproduction because of good environment and continue to the next migration.

At the reproduction phase, combining the merits of PSO algorithm (memory individual information) and the characteristics of biomes multiplication (population reproduction) makes the post-breeding population memory the mother possible. For example C1 reproduces two populations: C1' and C1", both of them will inherit the memory of C1 (memory includes the individual optimal location and current location of C1), and then migrate respectively to get C11 and C12.

4) Evolution

The reason why biological population is able to keep balance is not only the extinction of population but also the evolution of population. This constant evolution creates a lot of new population, which makes the whole system keep balance. This evolution is worth thinking, the phase of that is also contained in our algorithm. It makes the number of population hold the line, of course, more important; it will not become the monotonous population.

Mentioned above, it is said that there are *pm* populations to evolve, that is to say, it will creates *pm* new populations. However, we know that only the location can distinguish the differences in solution space. So *pm* populations evolve means generating randomly *pm* new solutions.

5) Algorithm Description

a) Initialize parameters and set the number of evolution population for pm;

b) Initialize population $X = (X1, X2, \ldots, Xn)$;

c) Calculate the fitness of each population;

d) Selection operation;

e) Reproduction and elimination operation;

f) Evolution operation;

g) Migration operation according to (2);

h) End if the migration algebra arrived; otherwise go to *c)*.

IV. EXPERIMENT AND RESULT

In order to verify the feasibility and effectiveness of the algorithm, consider the following two aspects. Among them, the feasibility of the algorithm is measured by the times converging to the optimal solution; while the effectiveness of the algorithm is measured by the average AVE of function value satisfying the iteration times and average time-consuming time.

A. Experiment 1

Choose the following functions testing the biologic population PSO:

$$f_1(x) = 0.5 - \frac{\sin^2\sqrt{x_1^2 + x_2^2} - 0.5}{[1 + 0.001(x_1^2 + x_2^2)]}, x \in [-10,10]$$

$$f_2(x) = \cos(5x_1)\cos(5x_2)e^{-0.001(x_1^2 + x_2^2)}, x \in [-2.048,2.048]$$

$$f_3(x) = (4 - 2.1x_1^2 + \frac{1}{3}x_1^4)x_1^2 + x_1x_2 + 4x_2^4 - 4x_2^2, x \in [-3,3]$$

f_1 has numerous local maxima points, only the point (0,0) for the global maximum 1. Around it there are a number of ridges whose peak changes gradually. The values near to the global optimal are all 0.990283 making it easy to stop at this local maximum point.

f_2 has more peaks in the definition region , where f (0,0) = 1 is the global maximum, the rest peaks are all quite near the highest point.

f_3 has six local minimum points, two global minimum points $f(-0.0898, 0.7126) = f(0.0898, -0.7126) = -1.031628$.

The parameters in PMPSO and PSO are set as follows: the greatest migration algebra $c1=c2=2$, $wmax=0.9$, $N=200$, $wmin=0.4$, $vmax=0.5$, $vmin=-0.5$; in PMPSO $pm= 5$; the number of population is 100.

Comparing PMPSO and PSO with DSGA of [9], get the following experimental results (repeat running 100 times as follows in Tab. 1. From Tab. 1 it can be seen that PMPSO is better than PSO and DSGA on the performance of global convergence. Moreover, reaching optimization 100% on testing f_1 and f_3 showed this algorithm is stable. Therefore, PMPSO is feasible on searching optimization.

B. Experiment 2

Choose the following functions:

$$f_4 = \frac{1}{4000}\sum_{i=1}^{30} x_i^2 - \prod_{i=1}^{30}\cos(\frac{x_i}{\sqrt{i}}) + 1, |x_i| \le 600$$

$$f_5 = \sum_{i=1}^{30}[x_i^2 - 10\cos(2\pi x_i) + 10], |x_i| \le 5.12$$

In which, f_4, f_5 get the minimum at (0, 0). The Parameter in PMPSO are set as follows: the largest migration algebra N = 2000, the number of population= 20, $wmax$ = 0.9, $wmin$ = 0.4, $vmax$ = 0.5, $vmin$ =-0.5, $c1$ = $c2$ = 2; in PMPSO pm = 2; at the same time, the condition of algorithm termination is f_4 <0.001, f_5 <0.001.

Comparing PMPSO with StPSO and StdGA in [9] get the following experimental results (the average running 100 times), illustrated as Tab. 2.

TABLE I. THE TIMES SEARCHING THE GLOBAL OPTIMAL BY EACH ALGORITHM

Function	PSO	DSGA	PMPSO
f_1	87	90	100
f_2	10	94	95
f_3	100	91	100

TABLE II. EXPERIMENTAL RESULTS

function		f_4	f_5
StPSO	AVE	0.0189±0.0586	49.4664±0.6299
	TIME	6.457s	5.890
StdGA	AVE	889.537±3.939	49.3212±1.1204
	TIME	20.350s	17.419s
PMPSO	AVE	0.00±0.00	0.00±0.00
	TIME	4.256s	4.136s

From Tab. 2 we can see that PMPSO is better than PSO and DSGA on the performance of the global convergence as well as the relatively less time. This shows the algorithm is effective.

V. CONCLUSIONS AND FUTURE WORK

This article introduces the survival of the fittest rules of biomes multiplication to PSO algorithm, eliminating inferior population and keeping superior population. It is helpful to make full use of population resources and speed up the algorithm convergence. At selection phase, the successful classification of the population by LEVEL SET theory makes the algorithm accord with the principle of survival of the fittest. At the same time, taking into account the evolution of population can make it keep diversity, which prevents the algorithm becoming monotonous and precocious. These new improvements enhance optimization accuracy and convergence speed of the traditional PSO as well as the capacity PSO algorithm solves complex problems. Finally, we verified with an example for the feasibility and effectiveness of the new algorithm PMPSO.

REFERENCES

[1] R C Eberhart and J Kennedy, "A New Optimizer Using Particles Swarm Theory," In: Proc Sixth International Symposium on Micro Machine and Human Science, Nagoya, Japan, 1995.

[2] Eberhart R C and Shi Y, "Particle swarm optimization: developments, applications and resources," Proc. Congress on Evolutionary Computation 2001. Piscataway, NJ:IEEE Press, pp. 81–86, 2001.

[3] Parsopouls K.E., Plagianakos V.P., Magoulas G.D., and Vrahatis M.N., "Stretching technique for obtaining global minimizers through particle swarm optimization," In: Proc Workshop on Particle Swarm Optimization, Indianapolis USA, pp. 22–29, 2001.

[4] Silva A., Neves A., and Costa E., "SAPPO: A simple, adaptable, predator prey optimizer," Lecture Notes in Artificial Intelligence, vol. 2909, pp.59–73, 2003.

[5] He S., Wu Q.H., Wen J.Y., Saunders J.R., and Paton R.C., "A particle swarm optimizer with passive congregation," BioSystems, vol. 78, pp. 135–147, 2004.

[6] Niu B., Zhu Y.L., and He X.X., "Construction of fuzzy models for dynamical systems using multiple cooperative particle swarm optimizer," Lecture Notes in Artificial Intelligence, vol. 3613, pp. 987–1000, 2005.

[7] Hendtlass T., "Preserving diversity in particle swarm optimization," Lecture Notes in Artificial Intelligence, vol. 2718, pp. 31–40, 2003.

[8] Li Qinghua, Yang Shida, and Yuan Youlin, "Improving Optimization for Genetic Algorithms Based on Level Set," Journal of Computer Research and Development, vol. 43, pp. 1624–1629, 2006.

[9] Liu Zhiming, Zhou Jiliu, and Chen Li, "A novel genetic mutation operator for maintaining diversity," Mini-Micro Systems, vol. 24, pp. 902–904, 2003.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 071-074

Optimal Operation of Hydropower Station by Using an Improved DE Algorithm

Lei Yin[1], and Xiaoxiang Liu[2]
[1] School of Mechano-Electronic Engineering Xidian University, Xi'an, China
Email: yinlei_w@163.com
[2] Department of Computer Science of Zhuhai College Jinan University, Zhuhai, China
Email: tlxx@jnu.edu.cn

***Abstract*—The mid-long term optimal operation of hydropower station is a nonlinear combinatorial optimization problem with strong constraint. This paper eliminates the problem constraints by introducing penalty function and solving this problem with DE algorithm, simultaneously, it proposes a few improvements according to the disadvantages of DE algorithm: adopt the scheme which generates differential vector to carry out mutation operation for optimum individual of each generation by randomly selecting four different individuals from population; and it adopts adaptive secondary mutation differential algorithm. The example demonstrates that DE algorithm is easy for programming realization and occupies little computer memory, and it is an effective search algorithm with rapid speed of convergence and superior precision as well as a new approach for multi-reservoir combined optimal operation to overcoming the curse of dimensionality.**

***Index Terms*—Differential evolution, hydropower station, optimal operation**

I. INTRODUCTION

The optimal operation of hydropower station is a complex combinatorial optimization problem. Classical optimal algorithms such as Dynamic Programming (DP), progressive optimization algorithm (POA) have their respective advantages as well as obviously deficiency when solving optimal operation of hydropower station problems. When solving with DP algorithm, as the number of hydropower station increases and the subdivision on optimization period, the calculation speed will decrease significantly, thus curse of dimensionality appears. If the hydropower station number is more than two, the occupancy computer memory of progressive optimization algorithm is thereupon increases, and calculation speed is significantly decreases [1].

In recent years, in the field of evolutionary computation, DE algorithm has attracted growing concern as an optimal algorithm based on swarm intelligence theory with excellent properties. It generates swarm intelligence for guiding in optimization search by cooperation and competition among individuals in the group. Compared with evolutionary algorithm, DE algorithm reserves global searching strategy based on population using Real-coded and differential based simple mutation operation as well as one to one competition survival strategy.

Compared with genetic algorithm [2], the most important feature of the differential evolution is using linear combination of parent-generation multiple-individuals in the generation process of each new individuals rather than the traditional and single chromosomal chiasma technology of genetic algorithm. Simultaneously, the own memory ability of DE algorithm makes it possible with dynamic tracking for the current searching conditions to adjust the search strategy, and it has relatively strong global convergence ability and robustness. This algorithm is not necessarily by means of the feature information of the problems and it is suitable for solving optimization problem in a complex environment to which mathematical programming method is not applicable. The application fields of DE is broader and broader which involves chemical, electric, mechanical design, economic, environmental engineering and operational research, etc [3].

This paper eliminates the optimal operation of hydropower station constraints by introducing penalty function and makes DE algorithm applicable to this problem. At the same time, it proposes a few improvements according to the disadvantages of DE algorithm: adopt the scheme which generates differential vector to carry out mutation operation for optimum individual of each generation by randomly selecting four different individuals from population; and it adopts adaptive secondary mutation differential algorithm. We solve the optimal operation of hydropower station model by improved differential algorithm, thus better results are expected to be obtained.

II. THE HYDROPOWER STATION OPTIMAL OPERATION MODEL

Suppose the research object is a multi-purpose reservoir, and the inflow runoff sequence of reservoir is known, the chosen object which satisfies the comprehensive utilization constraints maximizes the annual power generation income by considering the electricity price difference between wet season and dry season.

Objective function:

$$E = \max \sum_{t=1}^{T} A \bullet Q_t \bullet H_t \bullet M_t \bullet P_t \qquad (1)$$

corresponding author: tlxx@jnu.edu.cn

Constraints:

Water Balance Constraints:

$$V_{t+1} = V_t + q_t - Q_t - S_t \quad (2)$$

Reservoir Storage Constraints:

$$V_{t,\min} \le V_t \le V_{t,\max} \quad (3)$$

Discharge Volume Constraints:

$$Q_{t,\min} \le Q_t \le Q_{t,\max} \quad (4)$$

Power Station Output Constraints:

$$N_{\min} \le A \bullet Q_t \bullet H_t \le N_{\max} \quad (5)$$

In the formula: E is the annual power generation income, A is the coefficient of station comprehensive output; Q_t is the station generating flow in period t; H_t is the station average generation net head in period t; P_t is the station electricity price factor in period t; T is the annual calculation total intervals(the calculation interval is month, T=12); and M_t is the hours of the period t. V_{t+1} is the station end reservoir storage of period t; V_t is the station initial reservoir storage of period t; q_t is the average reservoir inflow in period t; Qt is the generating flow in period t; S_t is the abandoned water flow in period t. In the period t $V_{t,\min}$ is the minimum reservoir storage to be guaranteed; $V_{t,\max}$ is the maximum reservoir storage to be guaranteed. $Q_{t,\min}$ is the minimum discharge volume to be ensured, $Q_{t,\max}$ is the maximum discharge volume to be ensured. $N_{\min}$ is the minimum permissible station output; $N_{\max}$ is the limit of station maximum output (generally is installed capacity).

III. DIFFERENTIAL EVOLUTION TECHNIQUE

The basic operation of DE algorithm involves mutation, crossover and selection, but it is different from other evolutionary algorithm like genetic algorithm (GA). DE forms a population consists of N_p (population size) D-dimensional (variable number) parameter vector x_i^g (i=1, 2, …, N_p) to carry out optimization in the search space, where g represents the iteration times. Firstly, form mutation operator by the differential vectors among the individuals of parent-generation; then make crossover operation among the father-generation individuals and the variation individuals according to certain probability, thus generating an experimental individual. Afterwards, a selection operation is carried out based on the fit among the father-generation individuals and the experimental individuals, then choose individuals with a more superior fit as filial generation.

A. Mutation Operation

The basic mutation composition of DE algorithm is the differential vectors of parent-generation, each vector forms different differential evolution technique scheme for two different individuals (x^g_{r1}, x^g_{r2}) of parent-generation according to different generation method of mutation individuals. The conventional differential evolution technique usually randomly select three different individuals to form differential vectors, the formula of individual mutation operation is [3]:

$$x_m = x^g_{gbest} + F[(x^g_a - x^g_b)] + x^g_c \quad (6)$$

This paper randomly select four different individuals form population to carry on mutation operation for each optimal individual. This scheme can lift the convergence rate, and remains preferable population diversity to some extent, namely:

$$x_m = x^g_{gbest} + F[(x^g_a - x^g_b) + (x^g_c - x^g_d)] \quad (7)$$

In (6) and (7): x^g_{gbest} is the individual in the population with the best fit; x^g_a, x^g_b, x^g_c and x^g_d are x^g_{gbest} and four individuals which are different on one another; F is a scaling factor with a value range of (0, 1.2) which is equivalent to a noise version of x^g_{gbest}, the bigger F, the more mutation of x^g_{gbest}, and the greater impact on $x_{m.}$.

B. Crossover Operation

Crossover is the process of exchanging partial genes of two chromosomes and generating different chromosomes based on certain crossover probability CR. DE algorithm maintains population diversity by using crossover operation, for the number i individual x_i^g, a crossover operation is carried out with x_m to generate experimental individual x_T. To ensure the evolution of individual x_i^g, firstly, make at least one dimension in the D-dimensioned variables contributed from x_m by random selection, but for other dimensions,

we can determine which dimension of x_T is contributed by x_m and which dimension of x_T is contributed by x_i^g by making use of a crossover probability factor CR. The crossover operation formula is :

$$x_{Tj} = \begin{cases} x_{mj}, rand() \le CR \\ x_{ij}^g, \ rand() \ge CR \end{cases} \quad (8)$$

Where j=1, 2, …, d and rand() is the uniform random number with a value range of [0, 1], j represents the number j variable(gene), and D is the variate dimension D. Known from (8), the larger CR, the greater contributions from x_m to x_T, when CR=1,then x_m=x_T, and it is propitious to local search and the acceleration of convergence rate; while the smaller CR, the more contributions from x_i^g, when CR=0 then x_T= x_i^g and it is good for global search and maintaining population diversity. Thus, maintaining population diversity and convergence rate is contradictive.

C. Selection Operation

Selection is the process of choosing individuals with strong vitality in the population to generate a new population according to certain principles and methods based on the evaluation of the individual fit. The purpose of selection is to choose excellent individuals from the current population and provide them with the opportunities to multiply offspring for parent-generation. DE algorithm adopts "greedy" search strategy, and it generates experimental individuals x_T by mutation and crossover operation to compete with x^g_i. x_T is only selected as filial generation when the fit x^g_i is more superior, otherwise directly choose x^T_i as filial generation. The selection operation formula is

$$x_i^{g+1} = \begin{cases} x_T, & f(x_T) < f(x_i^g) \\ x_i^g, & f(x_T) \geq f(x_i^g) \end{cases} \quad (9)$$

D. Convergence Analysis of differential evolution

In (7), x^g_{gbest} is the current optimal solution, with the process of evolution, other individuals draw close to it rapidly. If x^g_{gbest} is a local optimal point, with the evolution of population, the difference between individuals becoming smaller, and mutation vector D_{ab} tends to 0, crossover and selection operation can not change population diversity, and all the individuals trend to x^g_{gbest}. Thus an anew search in the solution space is impossible. Therefore, the algorithm falls into local optimum and premature convergence phenomenon.

As differential evolution algorithm is a random search strategy, and realistic problems are always complex multi-peak, and has several local optimal points, thus algorithm is easy to fall into local optimum instead of finding global optimum. Even adopting

Any other individuals in the population instead of individuals of x^g_{gbest} which performs mutation can not avoid falling into local optimum, and convergence rate is slow down at the same time. Therefore, DE algorithm is easy to premature convergence and fall into local optimum [4].

To rapid the convergence rate and avoid premature convergence, people proposes several methods, where [5] puts forward adaptive secondary mutation differential evolution algorithm. This paper proposes a method of time-varying crossover probability factor *CR*, suppose CR_{min} is the minimum crossover probability, and CR_{max} is the maximum crossover probability, g is the current iteration times, G is the maximum iteration times, that is:

$$CR=CR_{min}+\frac{g(CR_{\max}-CR_{\min})}{G} \quad (10)$$

Thus making time-varying crossover probability factor *CR* linearly increased with the performing of algorithm iteration. x^g_i contributes the most to x_T in the initial stage, and the global search ability is improved, while x_m contributes the most to x_T in the later stage, and the local search ability is improved. This method rapid the convergence rate and improve the algorithm performance. Hence, this paper adopts this improvement and applies it to hydropower station optimal scheduling problem.

IV. ALGORITHM DESIGN

The optimal operation of hydropower station is a nonlinear and multistage combinatorial optimization problem with strong constraint, and it can be presented as: finding a water level change sequence which satisfies all constraints to maximize the annual power generation income. When solving the problem with DE algorithm, a individual is an operation strategy of hydropower station, and the element of individual position vector is the station final water level of each periods whose change should satisfy all constraints of the above mentioned model. To increase initial feasible solutions, this paper introduces penalty function to eliminate constraints [6]. The algorithm as follow:

1) Parameter Initialization: select population size N, weighted factor $F\in[0,2]$, maximum evolutionary generations $G_{\max}$, evolutionary generations G=0, hybrid rate $CR\in[0, 1]$;

2) Randomly generate an initial population: In the permissible water level change range of each period, randomly generate m group of final water level of each periods change sequence W^0. Where n is the period number, $W^0=\{w_i, 1\leq i\leq m\}$, $w_i=(a_1, a_2,\ldots,a_n)$,.

3) Carry on mutation operation based on (7);

4) Carry on crossover operation based on (8);

5) Carry on selection operation based on (9);

6) $G=G+1$,jump to *7)* if G is more than $G_{\max}$ or the precision meets the qualification, otherwise jump to *3)*;

7) Output results and operation time.

V. EXPERIMENT AND RESULT

To prove the feasibility and effectiveness of the above algorithm, a calculation is based on a certain reservoir example. Select the population size 40, and the maximum iteration times 100. This comprehensive utilized reservoir is mainly used for irrigation and water supplying as well as flood prevention and power generation, processing seasonal regulation ability. Known the hydropower station water level-- storage capacity, lower water level-- discharge volume, units presupposition output curve and the hydrograph of monthly average flow. The normal water level of reservoir is 877m , dead water level is 817m, force-voltage factor is 8.5, installing 760,000 kW, guaranteed output is 160,000 kW, the maximum flow through the turbine is $1000m^3\cdot s^{-1}$. It requires that the flood period (June-September) water level is less than limit level for flood control, namely 850.

Divide the period of hydropower station optimal operation into 12 intervals based on month, and discrete the reservoir water level into 60 states, the electricity price factor from December to April is 1.5, in November and May is 1, and it is 0.75 from June to October. The reference electricity price is 0.28 RMB/ kw·h. Solving the problem with dynamic programming and the above PSO respectively, see the results in Tab. 1 and Tab. 2:

TABLE I. DYNAMIC PROGRAMMING RESULTS

Time/month	Month end water level/m	reservoir inflow / $m^3\cdot s^{-1}$	generating flow / $m^3\cdot s^{-1}$	abandoned water flow / $m^3\cdot s^{-1}$	output/ten thousand kw
1	874.9	142	156	0	17.4
2	870.4	127	159	0	17.3
3	868.8	151	161	0	17.0
4	870.4	231	221	0	23.4
5	846.9	660	790	0	75.8
6	850	893	878	0	75.9
7	850	602	602	0	53.1
8	850	521	521	0	46.1
9	850	326	326	0	29.0
10	877	319	160	0	16.3
11	877	185	185	0	20.8
12	877	150	150	0	16.9
Annual income: 818,637,200RMB calculation time :3.15 s					

TABLE II. DE ALGORITHM RESULTS

Time/month	Month end water level /m	reservoir inflow/ $m^3 \cdot s^{-1}$	generating flow/ $m^3 \cdot s^{-1}$	abandoned water flow / $m^3 \cdot s^{-1}$	output/ten thousand kw
1	875.5	142	152	0	17
2	871.5	127	156	0	17.1
3	870.3	151	159	0	17.1
4	870.5	231	229	0	24.4
5	847.2	660	790	0	75.9
6	850	893	879	0	76
7	850	602	602	0	53.1
8	850	521	521	0	46.1
9	850	326	326	0	29
10	877	319	160	0	16.3
11	877	185	185	0	20.8
12	877	150	150	0	16.9
Annual Imcome:820,678,300RMB Calculation Time: 0.93s					

Through the process of analyzing optimal operation of hydropower station with DE algorithm and dynamic programming respectively, we find DE algorithm is superior to dynamic programming, and the annual power generation income can increase by 2041,100 RMB. DE algorithm has more superiority because it can adopt higher calculation precision (enlarge the population number). Simultaneously, when the storage capacity discrete points of reservoir is 60, 6012 state points should be saved in the computer using dynamic programming, and the huge data storage makes it difficult to solving optimal operation problem of hydropower station. However, saving these state points is not necessary when using DE algorithm and it greatly saves CPU time and memory demand quantity. DE algorithm is superior to dynamic programming in calculation speed and provides a new approach for multi-reservoir combined optimal operation to overcoming the curse of dimensionality.

VI. CONSTANT AND FUTURE WORK

This paper introduces the application of DE algorithm in the mid-long term optimal operation of hydropower station. The optimal operation of hydropower station is a nonlinear and multistage combinatorial optimization problem with strong constraint, and we eliminate the problem constraints by introducing penalty function and solving this problem with DE algorithm. Simultaneously, it proposes a few improvements according to the disadvantages of DE algorithm: adopt the scheme which generates differential vector to carry out mutation operation for optimum individual of each generation by randomly selecting four different individuals from population; and it adopts adaptive secondary mutation differential algorithm. The improved DE algorithm is applied to the optimal operation of hydropower station, and it has advantages of simple principles, easy for programming realization, small computer memory occupancy, rapid calculation speed, high search efficiency, and it provides a new approach for multi-reservoir combined optimal operation to overcoming the curse of dimensionality.

REFERENCES

[1] Zhang Shuanghu, Huang Qiang, and Sun Tingrong, “Study on the optimal operation of hydropower station based on parallel recombination simulated annealing algorithms,” Journal of Hydroelectric Engineering, vol. 23, pp. 16–20, 2004.

[2] Cheng R and Gen M, “Vehiele routing problem with fuzzy due-time using genetic algorithm,” Japanese Journal of Fuzzy Theory and Systems, vol. 7, pp. 1050–1061, 1995.

[3] Cao Eebao, Lai Mingyong, and Li Donghui, “Vehiele routing problem with fuzzy demands based on hybrid differential evolution,” Systems Engineering-Theory & Practice, vol. 29, pp. 106–111, 2009.

[4] Deng Zexi, Cao Dunqian, and Liu Xiaoji, “New differential evolution algorithm. Computer Engineering and Applications, vol. 44, pp. 40–42, 2008.

[5] Wu Lianghong, Wang Yaonan, and Yuan Xiaofang, “Differential Evolution Algorithm with Adaptive Second Mutation,” Control and Decision, vol. 21,pp. 898–902, 2006.

[6] S. Kannan, S. Mary Raja Slochanal, and P. Subbaraj, “Application of particle swarmoptimization technique and its variants to generation expansion planning problem,” Electric Power Systems Research, vol. 70, pp. 203–210, 2004.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 075-079

Combination of Cloud Model and Rough Set to Find Knowledge in IDSS for Intelligent Disaster Emergency Decision

Hongli Wang[1 2]
[1] School of Economy and Management, Zhongyuan University of Technology, Zhengzhou, 450007, China
[2] School of Management, Xi'an Jiaotong University, Xi'an, 710000, China
E-MAIL: graduated852@163.com

***Abstract*—Decision support system using data mining to find decision knowledge is called Intelligent Decision Support System(IDSS). Rough set as a data mining method commonly is used to find classification knowledge in IDSS. But the classic data mining based on rough set is short in dealing with the blank value data or data with the character of blurring and randomicity. Such data is called as imperfect data. In order to overcoming this shortcoming the method of combination of cloud model and rough set to find knowledge from imperfect data in IDSS is proposed. Firstly the cloud is used to depict the imperfect data by group decision. In the following, attribution generation based on cloud model is used to generate the upper concept layer. In this step the cloud model depicting the imperfect data is classified into the concept layer which is proximal to itself according to distance between two cloud models. Then rough set method is used to gain knowledge. Lastly an experiment is given to verify the validity of the method.**

***Index Terms*—Decision support system, data mining, cloud model, rough set**

I. INTRODUCTION

Intelligent decision support system(IDSS) based on data mining is the system in which data mining is used to gain decision knowledge. Rough set is used as a data mining method in IDSS. But data commonly is incomplete or uncertain in database in IDSS. Such data is called as imperfect data with the character of blurring and randomicity. Sometimes the data is blank value. But the blank value can be evaluated as the value of blurring and randomicity by experts. Imperfect data can not directly be input into neural network because of uncertain of value.

In order to overcome the defect of imperfect data some methods are proposed in relative literatures. They include fuzzy, rough set and grey theory and so on. For example the combination of fuzzy method and neural network is given to classify with incomplete data[1]. The classification method based on fuzzy method is argued and investigated[2]. Rough set and neural network is combined to predict software change[3]. Rough set and neural network is integrated to fault diagnosis[4]. Artificial neural network and grey systems is used for the prediction of slope stability[5]. Cloud model and neural network is used for induction of decision tree[6]. These methods have some shortcomings in the following. (1) Firstly the imperfect data is not evaluated by group experts for these methods. In fact man is not utterly ignorant the imperfect data but approximate cognition. But this cognition is difficult to be expressed by the certain manner. So they can be evaluated by group experts who master the domain knowledge on condition that approximate cognition can be expressed by more objective mathematics model and method. (2) Imperfect data is absent, uncertain or blank value. So the imperfect data is not be used to gain the knowledge. But data can be used to gain knowledge after pretreatment. In these methods the effective method of evaluation and attribution generation is not given to deal with imperfect data. (3) Expression of imperfect data is not objective and reasonable in some methods. Imperfect data is the data with the character of blurring and randomicity So more precisely expression of data should be use. In some methods only one aspect of character of data is expressed. Another aspect of character of data is ignored.

The method of combination of cloud model and rough set is proposed to find classification knowledge from imperfect data in intelligent decision support system in this article. In section 1 the background and general situation of problem is introduced. And the existing problem and general resolving scheme is proposed. In section 2 the method of cloud model and rough set is depicted. Cloud model is used as the pretreatment tool. Cloud is used as the mathematics model of group decision in the evaluation. Cloud is also used as the model of attribution generation. Then data is applied into rough set after the pretreatment. In section 3 the experiment is given to validate the validity of method of section 2. Lastly in section 4 the conclusion and appraisal is given.

II. METHOD OF COMBINATION OF CLOUD MODEL AND ROUGH SET

A. Data pretreatment based on cloud model

Data in database used to gain knowledge commonly is values of records. Every record is composed of the values of fields. The values of fields maybe have all kinds of types. These types include number value, qualitative characterization, imperfect data, range value, rank value and so on. All types except imperfect data may easily be dealing with into the number value.

Case of imperfectness of data mainly is in the following:

(1) The first case is that data on some fields for some records in the database is uncertain value, but the data can be evaluated as the value which is approximately about a point. Such value common is described as probably, maybe, assumably, in round numbers and so on because of blurring the subjective estimation. But other data on the fields for other records in the database is certain value. These values have the character of blurring and randomicity. In this case the value of fields can be expressed by the cloud model. The number feature of cloud is signified with the expected value, entropy and hyper entropy. The expected value, entropy and hyper entropy is gained by group decision which is made by group experts who master the relative information on that problem domain.

The process of group decision to gain the cloud model expression of fields is in the following. At first the evaluation language set is chosen. The evaluation language of single layer and evaluation language of multi-layer may all be used. The evaluation language set of single layer or multi-layer is built according to the actual problem. The sketch maps of evaluation language set of single layer and multi-layer are given in Fig.1.

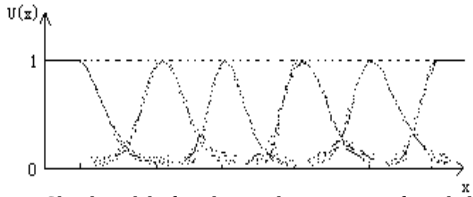


Fig.1 Cloud model of evaluation language set of single layer

The expected value, entropy and hyper entropy of cloud model are gained by the following method[7,8]. The approximate value range of fields $[D_{min}, D_{max}]$ is ascertained by expert. N clouds is created to denote such a set of cloud models such as (approximate …, approximate …,…, approximate …). The center cloud model is recorded as $A_0(Ex_0, En_0, He_0)$. The left and right cloud models are recorded as $A_{-1}(Ex_{-1}, En_{-1}, He_{-1})$, $A_{+1}(Ex_{+1}, En_{+1}, He_{+1})$, $A_{-2}(Ex_{-2}, En_{-2}, He_{-2})$, $A_{+2}(Ex_{+2}, En_{+2}, He_{+2})$, …, $A_{\frac{n-1}{2}}(Ex_{\frac{n-1}{2}}, En_{\frac{n-1}{2}}, He_{\frac{n-1}{2}})$, $A_{+\frac{n-1}{2}}(Ex_{+\frac{n-1}{2}}, En_{+\frac{n-1}{2}}, He_{+\frac{n-1}{2}})$ (n is odd number). The expected value, entropy and hyper entropy of cloud model are gained by the golden section method. Hypothesis n=5, then expected value, entropy and hyper entropy of cloud model are calculated by the golden section method in the following formulas[7]:

$$Ex_0 = (D_{min} + D_{max}) / 2 \quad (1)$$

$$Ex_{-2} = D_{min} \quad (2)$$

$$Ex_{+2} = D_{max} \quad (3)$$

$$Ex_{-1} = Ex_0 - 0.382*(D_{max} - D_{min}) / 2 \quad (4)$$

$$Ex_{+1} = Ex_0 + 0.382*(D_{max} - D_{min}) / 2 \quad (5)$$

$$En_{-1} = En_{+1} = 0.382*(D_{max} - D_{min}) / 2 \quad (6)$$

$$En_0 = 0.618\, En_{+1} \quad (7)$$

$$En_{-2} = En_{+2} = En_{+1} / 0.618 \quad (8)$$

Given He_0, then

$$He_{-1} = He_{+1} = He_0 / 0.618 \quad (9)$$

$$He_{-2} = He_{+2} = He_{+1} / 0.618 \quad (10)$$

By this step, language sets and their cloud models is created. The nature language set is used to evaluate the value of fields by the experts. Then the number feature of cloud model of nature language given by expert is aggregated to gain the integrated evaluation value of every field. The aggregate formula is in the following.

$$Ex' = \beta_1 Ex_1 + \beta_2 Ex_2 \quad (11)$$

$$En' = \frac{En_1(Ex_2 - Ex) + En_2(Ex - Ex_1)}{Ex_2 - Ex_1} \quad (12)$$

$$He' = \frac{He_1(Ex_2 - Ex) + He_2(Ex - Ex_1)}{Ex_2 - Ex_1} \quad (13)$$

Thereinto $A(Ex', En', He')$ is the cloud model aggregated. $A(Ex_1, En_1, He_1)$ or $A(Ex_2, En_2, He_2)$ is the model which is given by the expert. β_i (i=1, 2) is the aggregate coefficient. Commonly value of β_i (i=1, 2) is 0.5. β_i (i=1, 2) can be adjusted according to factual instance.

B. Attribution generation based on cloud model

The field in database is also called as attribute. Attribution generation is defined as producing high-level concept layer according to the low-level concept layer or value of attribute. Other data in fields on the records is normal. They are disposed by attribution generation based on cloud transform method and classification. The non-regular data distributions are transformed to the overlaid cloud model by mathematical method of cloud transform[9,10]. Cloud transform method is used to produce the inter-overlay basic cloud model. At first the data diagram of column shape is created according to the database(Fig.2).

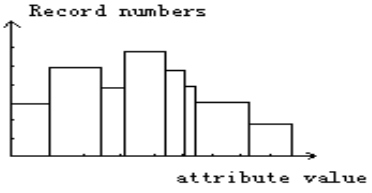


Fig.2 the data diagram of column shape in database

Then cloud transform method is used to produce the inter-overlay basic cloud model according to the data diagram of column shape. The formula of cloud transform is the following[8]:

$$f_i = \sum_{j=1}^{m} C_j Trap(i, Ex_{1_j}, Ex_{2_j}, En_{1_j}, En_{2_j}, He_{1_j}, He_{2_j}) + \varepsilon_i \quad (14)$$

Thereinto f_i is the value of function in the position i, ε_i is the surplus, m is the number of cloud.

The course of cloud transform is to find out the expected value、entropy and hyper entropy of every cloud. This method reflects the distribution of data in the domain while keeping the soft boundaries. The concept domain can be divided many zones expressed by the cloud model.

The value domain of field is dealt with into nature language expressed by model cloud. These cloud model are recorded as $A(Ex_1, En_1, He_1)$, $A(Ex_2, En_2, He_2)$,…, $A(Ex_n, En_n, He_n)$ called as the cloud model of attribute. They respectively denote the concept of (approximate …, approximate …,…, approximate …).

Then classification is used to classify the value of every field into the cloud model of attribute $A(Ex_1, En_1, He_1)$, $A(Ex_2, En_2, He_2)$,…, $A(Ex_n, En_n, He_n)$. The value of every field on the record in database is input the X condition cloud generator of every cloud model. The sketch map of X condition cloud generator is in the following in Fig.3[9,10]:

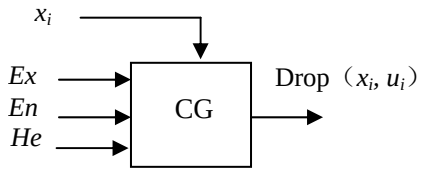


Fig. 3 X condition cloud generator

Hereinto x_i denotes the value of every field . u_i denotes the degree of subjection.

Then the degree of subjection is output by X condition cloud generator. The maximal degree of subjection and corresponding cloud model are recorded. The value of every field is classified into the corresponding cloud model of maximal degree of subjection. The value of fields is replaced by nature language of the corresponding cloud model in the database.

Lastly the cloud model of evaluation of imperfectness of data base on group decision $A(Ex', En', He')$ is classified into the cloud model according to the distance between $A(Ex', En', He')$ and the cloud model of attribute $A(Ex_1, En_1, He_1)$, $A(Ex_2, En_2, He_2)$,…, $A(Ex_n, En_n, He_n)$. The distance is calculated by the following formula (16)-(18):

$$D_{1,2} = \left| Ex_1 - Ex_2 \right| \qquad (15)$$

$D_{1,2}$ denotes relative distance between two cloud model.

$$FD_{1,2} = \left| En_1 - En_2 \right| \qquad (16)$$

$FD_{1,2}$ denotes relative distance of bluring between two cloud models.

$$RD_{1,2} = \left| He_1 - He_2 \right| \qquad (17)$$

$RD_{1,2}$ denotes relative distance of randomicity between two cloud model.

If all D are different $A(Ex', En', He')$ is classified into the cloud model of attribute in which D is minimal. If all D are same $A(Ex', En', He')$ is classified into the cloud model of attribute in which FD is minimal. If all D and FD are same $A(Ex', En', He')$ is classified into the cloud model of attribute in which RD is minimal.

C. Knowledge induction using rough set[11,12,13]

Rough sets have been introduced as a tool to deal with inexact, uncertain or vague knowledge in artificial intelligence applications. We recall some basic notions related to information systems and rough sets. An information system is a pair A = (U, A), where U is a non-empty, finite set called the universe and A - a non-empty, finite set of attributes, i.e. a: $U \to V_a$ for $a \in A$, where V_a is called the value set of a. Elements of U are called objects and interpreted as, for example, cases, states, processes, patients, observations. Attributes are interpreted as features, variables, characteristic conditions, etc.Every information system A = (U, A) and non-empty set $B \subseteq A$ determine a B-information function:
$Inf_B: U \to P(B \times \bigcup_{a \in B} V_a)$ defined by $Inf_B(x) = \{(a, a(x)): a \in B\}$.

We define B - indiscernibility relation as follows: $xIND(B)y$ if $Inf_B(x) = Inf_B(y)$.

For every subset $X \subseteq U$ we define the lower approximation $\underline{IND(B)}(X)$ and the upper approximation $\overline{IND(B)}(X)$ as follows:

$$\underline{IND(B)}(X) = \{x \in U : [x]_B \subseteq X\} \qquad (18)$$

$$\overline{IND(B)}(X) = \{x \in U : [x]_B \cap X \neq \varnothing\} \qquad (19)$$

We consider a special case of information systems called decision tables. A decision table is any information system of the form $\boldsymbol{A} = (U, A \cup \{d\})$, where $d \notin A$ is a distinguished attribute called *decision*. The elements of A are called *conditions*.

One can interpret a decision attribute as a kind of classification of the universe of objects given by an expert, decision-maker, operator, physician, etc. The cardinality of the image $d(U) = \{k: d(x)=k \text{ for some } x \in U\}$ is called the rank of d and is denoted by $r(d)$. We assume that the set V_d of values of the decision d is equal to $\{1,...,r(d)\}$. Let us observe that the decision d determines the partition $CLASS_A(d) = \{X_1,...,X_{r(d)}\}$ of the universe U, where $X_k = \{x \in U: d(x)=k\}$ for $1 \le k \le r(d)$. $CLASS_A(d)$ will be called *the classification of objects in* $\boldsymbol{A}$ *determined by the decision* d. The set X_k is called the k-*th decision class of* $\boldsymbol{A}$. The set $POS(B,\{d\})$ is called the positive region of classification $CLASS_A(d)$ and is equal to the union of all lower approximations of decision classes.

D. The combination of cloud model and rough set method

After data pretreatment based on model data may be input into the neural network. The step of combination of cloud model and neural network method is the

following:

Step1: Data in database is pretreated using the method in section 2.1. After that the value of fields is depicted by the cloud model expressed by natural language.

Step2: The different cloud model depicting the value of fields in the database is numbered as 1,2,…,n(n is integer). Here the same cloud model depicting the value of fields in database is numbered as the same number. 1,2,…,n is used to replace the corresponding cloud model depicting the value of fields in database.

Step3: The data is dealt with the rough set to find knowledge in the data table.

Ⅲ THE EXPERIMENT OF GAINING KNOWLEDGE IN IDSS

The availability of method is examined by the following experiment for disaster emergency decision. The data of degree of flood disaster in the database are given in table Ⅰ. The attribute of rank of degree is the decision attribute. It is regarded as the decision attribute of rough set.

The blank values in database are the imperfect data. They are gained by the group experts' evaluation using the method in section2.1.2.

TABLE Ⅰ.
DATA OF QUALITY RANK OF COTTON WOOL IN DATABASE

No.	Economy lost	Relative accident	Animal Death mumber	intension degree	plan damage mumber	Man Death mumber	Water deepth	Flood Area	measure taken
1	345	26	1643	50	120	134	20.25	11132	3
2	768	55	987	69	230	264	87.17	88156	1
3	667	17	1123	22	160	37	55.82	6480	4
4	213	29		21	653	36	35.95	7102	5
5		35	2567	43	321	77	74.91	90144	4
6	631	33	1896	32	115	52	28.67		2
…	…	…	…	…	…	…	…	…	…
200	879	67	3012	87	1370	98	79.12	99121	1

TABLE Ⅱ.
DATA OF QUALITY RANK OF COTTON WOOL IN DATABASE AFTER PRETREATMENT

No.	Economy lost	Relative accident	Animal Death mumber	intension degree	plan damage mumber	Man Death mumber	Water Deepth	Flood Area	measure taken
1	2	3	2	3	1	5	2	2	3
2	5	5	1	4	2	5	7	7	1
3	4	1	1	2	1	1	5	1	5
4	1	3	3	2	5	1	3	1	5
5	2	4	3	3	3	2	7	7	4
6	4	3	2	3	1	2	2	6	2
…	…	…	…	…	…	…	…	…	…
200	6	6	5	7	7	4	7	7	1

TABLE Ⅲ.
THE REGULATION INDUCED BY ROUGH SET

No.	Man Death mumber	Relative accident	intension degree	Water Deepth	Flood Area	measure taken
1	5	3	3			3
2				7	7	1
3	1		2		1	5
4		4		7	7	4
5		3		2	6	2

After evaluation the blank valuation is replaced by cloud model. Other values of every field in database are dealt with by attribution generation based on cloud model in section2.1.3. The hint figure of other data of attribute "intension degree" is dealt with in the following(Fig.4). There are seven cloud models in the Fig.4. They respectively

denote lowest, lower, mid-low, middle, mid-high, higher, highest. All data of attribute "impurity degree" should be classified into the corresponding cloud model of these models according to the value of u_i in Fig.5. After that the value of field in database is depicted by cloud model. Cloud model of imperfectness data is classified into the corresponding cloud model using method in section2.3.

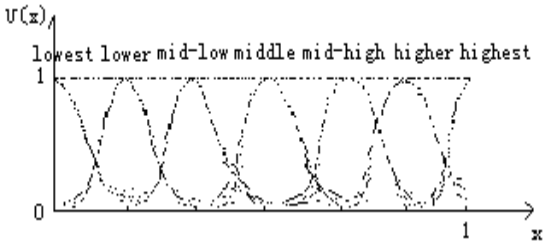


Fig.4 The attribution generation of attribute "impurity degree" on the No.5 record

The method of dealing with other data is the same to above method and step.

Then these cloud models are numbered by nature number. These numbers are used to replace the corresponding cloud model in database. Another method is that the maximal subjection degree is used to replace the corresponding cloud model. The former is used here. When there are two records are same except the attribute of "No.". They are viewed as two records. The result is shown in table Ⅱ. Then the data in table Ⅱ is applied the rough set to gain the decision knowledge for emergency decision. The following decision regulations are induced from the decision table by rough set(see table Ⅲ).

Ⅳ CONCLUSION

In order to deal with the incomplete or uncertain data in database group evaluation based on cloud model and the cloud transform method combined with rough set is proposed to find classification knowledge in intelligent decision support system above. Group decision based on cloud model used as the method of evaluation of imperfect data is impartial method. This method approximately furthest reflects the value of imperfect data. Then the attribution generation of data is executed based on cloud model. Then rough set is used to train and gain the classification knowledge. At last the experiment is given to validate the validity of method. The experiment manifests more precise classification knowledge can be gained by this method from imperfect data.

ACKNOWLEDGEMENTS

This paper is sponsored by government decision project of tendering & bidding of Henan province.

REFERENCES

[1] Chee Peng Lim, Mei Ming Kuan and Robert F. Harrison Application of fuzzy ARTMAP and fuzzy c-means clustering to pattern classification with incomplete data. Neural Computing & Applications. Vol 14,No.2, 2005. pp.104-113

[2] Yi-Chung Hu. Fuzzy integral-based perception for two-class pattern classification problems.Information Sciences.vol177,No.7, 2007. pp.1673-1686

[3] Sheela Ramanna. Rough Neural Network for Software Change Prediction. Lecture Notes in Computer Science.Vol 2475,2002. pp.602-609

[4] Qingmin Zhou and Chenbo Yin. An Integrated Approach to Fault Diagnosis Based on Variable Precision Rough Set and Neural Network. Lecture Notes in Computer Science. Vol 3498,2005.pp.514-520

[5] P. Lu and M. S. Rosenbaum Artificial Neural Network and Grey Systems for the Prediction of Slope Stability. Natural Hazards. Vol 30,No.3,2003. pp.183-398

[6] Yongqing Tian, Guoning Du, Zhi Li. Zhong-ying Zhu. Induction of Decision Trees Based on a Cloud Model Neural Network. Journal of Shanghai Jiaotong University. Vol 37,2003.pp.113-117.

[7] Chen H, Li D. Normal cloud model and its applications in KDD. Journal of Institute of Communications Engineering, , vol12,No4, pp. 39-44, 1998 (in Chinese)

[8] Fan J, Li D. Mining classifications knowledge based on cloud models. In: Proceeding soft the 3rd Pacific-Asia Conference On Knowledge Discovery & Data Mining, Beijing, China, pp.26-28,1999.

[9] Kai-chang Di, De-yi Li, De-ren Li. Cloud theory and its applications in spatial data mining and knowledge discovery. Journal of Image and Graphics,vol4, No.11,pp.930-935. 1999

[10] Deyi Li, KaiChang Di, Deren Li. Mining association with Linguistic cloud models. In Proceeding soft the Second Pacific-Asia Conference on Knowledge Discovery & Data Mining Melibourn, Australia, pp. 392-394,1998.

[11] Pawlak Z. 1991. Rough Sets. Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, 1991.

[12] Skowron A., Stepaniuk J. 1995. Generalized Approximation Spaces, In: Soft Computing, T.Y.Lin, A.M.Wildberger (eds.), San Diego Simulation Councils, Inc., 1995, pp. 18-21.

[13] Stepaniuk J. 1996. Similarity Based Rough Sets and Learning, Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery, November 6-8, 1996, Tokyo, Japan, pp. 18-22

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 080-083

Research into Models of Intelligence - type Multimedia Teaching Software Focusing on Thinking

Lan Wang
Computer and Information Engineering Institute
TianJin Normal University, TJNU
Tianjin, China
tjnuwanglan@163.com

***Abstract*—This article proposed the models of intelligence - type multimedia teaching software focusing on thinking. The design of the models reflects the characteristics of network , intelligence , multimedia , the design of modules and focusing on thinking. It also constructed a platformof developing teaching softwares , on which the brand - new I-CAI teaching softwares can be developed and the existent CAI teaching softwares can be completely applied.**

***Index Terms*—focusing on thinking ; intelligence - type ; multimedia ; development platform**

I. INTRODUCTION

With the proliferation of computers and the Internet's development, CAI software will become more extensive, more in-depth and more successfully applied to schools and families. CAI in the near future will become the main mode of global education. Education in the future will be a breakthrough in traditional education resources, geography and time constraints, is a student-centered education model of globalization. This model calls for CAI system of education should at least have the following conditions: network-based, that is, to the Internet as the next big virtual classroom education, network-based education are to achieve the basic conditions of globalization; intelligent, student-centered CAI software, must be can reflect the abilities and be able to target each student characteristics reflect the strong adaptability, intelligent teaching are the key to realize globalization; multimedia, CAI should make full use of computer technology, in many ways the vivid display of knowledge, to achieve The purpose of recreation; reusability, constructing an open software development platform that can ICAI to develop new teaching software, but also can have a little of CAI teaching software modifications, embedded into the system, to become part of the system to better share resources, so that it can become a CAI users technical support. At the same time has not yet found a good embodiment of this four goals CAI software, most of CAI software is just a simple display software, courseware, and its teaching model is based courseware as the center, rather than student-centered. Therefore, in view of these circumstances, put forward ideas based on reusable intelligent multimedia teaching software model, allowing the system to collect network-based, intelligent and multi-media and reusability in one.

The thesis is supported by Tianjin Municipal Educational Commission.It is "The network teaching platform in the automatic acquisition of user knowledge based on data mining ".

II. THE OVERALL STRUCTURE OF THE SYSTEM

Design reuse based on thinking of smart multi-media teaching software model, the system will be teaching content and teaching strategies separately, through the student model and individual guidance rules, suitable for dynamically generated content individualized teaching; by tracking students studying the situation, at any time to update Record students to realize the effect of individualized teaching. Therefore, in system design to individual, intelligence, reusability as the system design objectives. System architecture as shown in Figure 1.

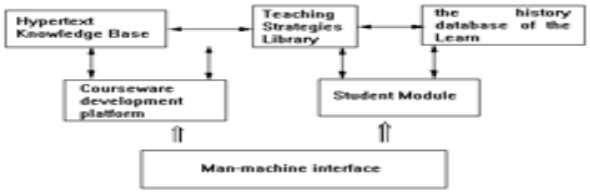


Figure1. System architecture

The system flow for the job: from "Student module" of students will enter the names of the students log on to the student file, and enter the student study unit (s) to find the corresponding "study the history of the Treasury," then "software development platform," according to study unit, students study the situation of (the beginning of none), hypertext knowledge base and teaching strategies to generate the best teaching sequence, and began teaching, at the same time teaching the rules in accordance with the guidance given to students, when students can conduct an interactive question and answer study . After the study, "Teaching Strategies Library" in the teaching module to calculate score keeping students studying library history, teaching strategies Treasury in accordance with the learning of the students on the knowledge base automatically adjust the relevant parameters, and so the cycle of teaching until the end of the event.

III. SYSTEM FUNCTION

A. *Hypertext Knowledge Database*

System Hypertext Knowledge is the core of the whole system, are combined with artificial intelligence and database product, is a knowledge representation, logical reasoning and data retrieval of knowledge in one treatment system, and substantial knowledge of information material provided to teachers and students in classroom teaching free to use. Hypertext-based multimedia knowledge base model, each knowledge point is a hypertext document, the link between knowledge points are hyperlinks, students and teaching strategies module library to form stored in relational databases, such as ORACLE, such as SYBASE and ACCESS. The advantages of this model are: knowledge of web publishing point is simple and convenient, and can easily manage multi-media information. Students set up the use of relational database modules and teaching strategies database, students can reflect the information and be able to perform simple reasoning.

- Knowledge Database Type

(1) multi-media teaching database. This is the point of basic knowledge, according to certain rules of search and classified information on organization of material, including graphics, text, sound, animation and video material, such as multi-dimensional information resource library. Authorware multimedia authoring system for users in recent years are a complete tool for the development of multimedia material. It not only provided with the application of standard interfaces to support dynamic link libraries, but also can run on its procedures interner.

(2) Micro-teaching modules library. Micro-teaching modules to help teachers in the teaching is a difficult teaching or to help students study for the purpose of a certain knowledge and skills designed to "small courseware." Its design and development methods and the original courseware for similar and different is that it generally do not need to cover design, no superfluous background, explanation and dubbing. Micro-teaching unit volume dapper, in line with the requirements of courseware platform combinations to facilitate retrieval and group access to use in teaching situations. In addition, the courseware has been "courseware development platform" after processing can be classified as such "small courseware."

(3) Virtual courseware resource library. Taking into account the current global network of trend of development, a school, a regional, national and even the teaching of information resources all over the world, can by teachers in classroom teaching retrieval, reorganization, combined with the flexibility to use current teaching needs.

- Knowledge database of the structure and relations link

Knowledge of teaching modules may be considered in two tiers: the concept of layer and physical layer, the concept of a knowledge-point layer nodes according to their mutually supportive relationship link from reticular structure, physical layer is based on the page for the node link into branch line and reticular formation. System to the physical layer access to knowledge through the concept of layers to carry out. Among them, the knowledge points each composed by a number of pages.

(1) the composition of knowledge points. To express the area of the intrinsic link between knowledge, domain knowledge can be divided into several chapters according to units, each unit is divided into a number of knowledge points, in order to dynamically adjust the teaching sequence, to provide guidance information, in the knowledge of the existence of points can be set up degrees , domain values, such as information, knowledge node information is divided into two parts: the index part and description part. Index part of the instructions how the system from the physical layer and look for information in the linked list does not include the information itself. A knowledge of the data model, including node identifier, the existence of degrees, through the threshold, skip threshold, the node body, chain index. Among them, the existence of degrees express the importance of knowledge points, giving it a certain range, for example, one <the existence of degree <8; express through the threshold only when knowledge point scoring not less than this threshold only allowed to leave the knowledge points , points into the next study of knowledge; skip threshold express knowledge points when the score is greater than the threshold, you can skip to the knowledge of points, points directly into the next study of knowledge; node pointer express body not because of knowledge points including the teaching of specific content knowledge, and its specific contents are several pages to describe and, therefore, the node pointer to the body of knowledge points, corresponding to the collection page.

(2) chain. Knowledge because knowledge is the definition of teaching, so the relationship between teaching knowledge can be manifested in the concept of layers. The relationship between knowledge points are complex. To this end, the system can be defined to support the chain, between the relevant knowledge that a mutually supportive relationship, and express the knowledge intensity of one point of contact between the degree of support. Supporting chain point to point direct support of the knowledge of knowledge point set, they will point to connect the knowledge into network structure. In addition, knowledge can also be defined between the best path chain, it can in supporting the chain linking into the network structure on the reasoning, the best path dynamically generated links, and to provide guidance to students. The best path chain teaching strategies can be dynamically generated and modify the library.

(3) Knowledge of the concept of layers and set up links with the physical layer. Knowledge of the concept of layers and set up links with the physical layer network structure mainly through the realization of the start-up module. Between knowledge because of their mutually supportive relationship are connected into a network structure, and the knowledge of points does not include the teaching of specific content knowledge and their specific contents are several pages to describe the. Therefore, in the start-up module, in addition asked to enter a description of each knowledge node information, but also asked to enter their corresponding page collection of the first page identifier (ie, icon names) and each Supporting information chain. Knowledge of which enter each node corresponding to the page set identifier of the first page are used for the purpose of establishing the concept of layer and physical layer link, and enter each chain Supporting information will be used for set up between the concept of knowledge layers the network structure. Therefore, when the information input after the completion of the module that is the whole concept of hypertext knowledge base layer and set up links with the physical layer. The module can be used when the realization of the adjacent table as the concept of Knowledge storage layer structure, may consider the use of Visual C + ten program, and generate a dynamic link library in Authorware call. Therefore, in Authorware courseware can be a teaching unit in each of the menu interface designed to initialize the button in the button to call the dynamic link library initialization function to complete these functions.

B. Teaching Strategies Library

Teaching strategies are manipulated to a certain degree of teaching content to students in the form of reasoning agencies, according to the students the content of modules and Knowledge to make intelligent decisions and to complete intelligent navigation. It can be understood as starting from the teaching objectives, according to students, the design and adjust the system of teaching sequences.

Teaching strategies with the best storage library path, teaching strategies and teaching the rules, teaching modules and the network structure, such as start-up module. Classroom teaching methods and strategies for hundreds of thousands, but the teachers are the most commonly used method of several. Such as on the way, quiz mode, exercise training mode, memory mode and hands-recited methods, the way different strategies will be designed to be filled with the reorganization of the framework in order to express simple icon, so that teachers in teaching according to their own needs different materials, micro-teaching modules with different ways of combining teaching strategies in order to flexibly cope with a variety of teaching situations.

C. Student Module

Student module records the students knowledge level, students are a reflection of the knowledge structure. Generally speaking, students have learned are from

Domain knowledge, so much a part of the organization by way of the field of Knowledge students set up study library. Knowledge Base because of the area is divided into chapters by a number of teaching modules, each module was divided into a number of knowledge points, so each module students have the corresponding libraries, used to record student chapter of the study, the students study database composed by the Record, the Record and in the teaching module knowledge points-one correspondence. In general, when students take part in a certain unit tests, the system will score the student's knowledge base to automatically adjust the internal structure of the concept of layers, such as the existence of degrees, contact intensity, making the teaching model can be dynamically according to teaching strategies, knowledge base and student model to generate the best instruction sequence, to realize individualized teaching purposes. The module can use data mining techniques to realize. Data mining, also known as data mining, are in accordance with the established targets from a large amount of data extracted and can be an effective person to understand the process, and ultimately the basis of prediction of data extraction and analysis. ICAI System in the prediction and analysis of data, mathematical models should be consistent with the epistemology and pedagogy, psychology and so on.

D. Courseware development platform

System courseware development platform have the following functions: (1) automatic generation of courseware guide. According to study unit, students study the situation, Knowledge hypertext network structure and teaching strategies to generate the best teaching sequence. Can also guide teachers every step of the operation, automatic generation of courseware. (2) input and output of "small software" function. Can be a complete dismantling of the courseware as a separate micro-teaching modules, but also be able to quickly and multiple micro-teaching modules and relevant content in the knowledge base into a courseware. (3) multimedia authoring features. Supports a variety of popular media material (swf, rm, etc.); rich interactive way. (4) to generate product features. Courseware can generate an executable file or html documents to meet the stand-alone or network environment screening requirements.

IV. CONCLUSION

The system design reflects the network-based, intelligent, multimedia-based, modular design and reusable ideas. Consistent with student-centered education model of globalization and the development direction of software engineering. As a new development of CAI teaching system, the current theoretical research, whether or application development is no doubt still immature side, there are many technical issues need to be further studies, such as Hypertext Knowledge Base database application in object-oriented database, connection information you can use an array to store, in the algorithm can be used directly, without having to do the conversion job. In this way, makes the description of knowledge points a more comprehensive and accurate. With the trust of the system will facilitate more sophisticated teachers to start their

own production of courseware, application software, computer-aided teaching techniques to promote a more in-depth, extensive and effectively applied to teaching activities.

REFERENCES

[1]http://www.etr.com.cn/
[2]http://211.152.9.125/
[3]http://disted.tamu.edu/edtclink.htm
[4]http://carbon.cudenver.edu/~lsherry/pubs/issues.htm

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 084-087

The Research of Intelligent Tutoring System Based on the Cognition and Emotion

Lan Wang
Computer and Information Engineering Institute
TianJin Normal University, TJNU
Tianjin, China
tjnuwanglan@163.com

***Abstract*—The thesis is supported by the doctoral fund Item of Tianjin Normal University. It is "Emotional and Cognitive Double Coding in Network Teaching Platform". This thesis aims at the research of Cognition and Emotion in Intelligent Tutoring System, and how to apply the Cognitive and Emotional in the "Intelligent Navigation Network Teaching Platform and Learning System". We can realize the individual study of Network Intelligent Tutoring System in both Cognitive and Emotional aspects.**

***Index Terms*—intelligent tutoring system, cognitive and emotional double coding, model students**

I. INTRODUCTION

It can achieve individuational education with a cognitive level in traditional network intelligent tutoring system, but it can not be perceived emotion of students in the process of teaching, it can not be achieved emotional aspects of individuational education. The main topics of this item is the key of the application to the cognitive theory and emotional theory, we update the building method about the student's learning model and teaching strategy in intelligent navigational learning system in order to be a great intelligent network teaching platform with cognitive and emotional interaction.

II. THE STUDENT'S MODULE BASED ON COGNIZE AND EMOTION

It is the key to intelligent tutoring system that Student's module should reflect the level of students to master the knowledge. In this paper, we try to build the student's medule combining with the emotion, psychology, ability, knowledge, interest, and so on, It can fully reflect the personality of the students. At present, students model focused on student behavior uncertainty reasoning, which is the key to break the above difficult. In this paper, we give a method of reasoning ability and easier to achieve with a fuzzy combination of theory, covering model thinking. The student model of the system include the following: 1. The student's module based on cognize and emotion; 2. The module about Student's cognitive state; 3. The module about Student's Emotional state; 4. The module about Student's study procedure; 5. The module about Student's study interest.

The thesis is supported by the doctoral fund Item of Tianjin Normal University.It is "Emotional and Cognitive Double Coding in Network Teaching Platform ".

AP-PROC-CS-09CN005

III. THE IMPLEMENT METHOD OF STUDENT'S MODULE

A. *The expression method in the Student's module*

- The expression about Student's cognitive state

The level of knowledge is used for each student to master knowledge points. In this paper, the student's cognitive level is divided into 6 grades 1,2,3,4,5,6, each attached level in the 6 grades is $\mu_K(i), i = 1 \sim 6$, we use K to express the fuzzy sets to the student's knowledge level.

$$K=\sum_{i=1}^{6} \frac{\mu_K(i)}{i},$$

$$0 \leqslant \mu_K(i) \leqslant 1, \quad \sum_{i=1}^{6} \mu_K(i) \leq 1$$

- The expression about Student's Emotional state

This is the relation about the students and learning emotions in Figure 1. Learners was attracted by a point of new knowledge,then this can generate interest for him in learning, or learners was confused by a new point of knowledge,then this can have been inspired by a strong desire for him to learn. Related to the using of algorithms to track changes in mood, the mood changes learners effective monitoring and analysis, forecasting and learners change trend sentiment. This study of students in a timely manner to carry out crisis intervention and adjustment, as well as the use of data mining algorithms to the learner's individual learning styles and learning for the acquisition, given its personalized learning.

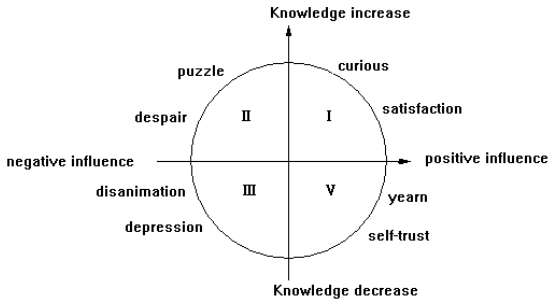


Figure 1:the relation about the students and learning emotions

- The expression about Student's study interest

For students, there are a lot of interests.In this article only be helpful to learn their preferences, the system gives the two types of interest: the presentation of knowledge and the way of teaching of knowledge. Knowledge of the presentation refers to how the media to show the form of learning content; education is the way the students like the manner in which access to education. Their value as shown in the next Table:

TABLE Ⅰ. Possible Value of Interests

interests	possible value
Knowledge of the presentation	text, pictures, slides, video, audio, virtual environment
Knowledge of teaching methods	Abduction method, deduction method, explore method

- Use a zero before decimal points: "0.25", not ".25". Use "cm3", not "cc". (*bullet list*)

B. The module about student's basic information

Students basic information module reflect the basic situation of individual students. Initial information is that the students first log on, the system through the interface their basic registration information. According to the relative stability of the information is divided into different static and dynamic information, including the study of static information, name, gender, professional faculties, such as relatively static information in the future in the process of learning can be modified by the students ; And refers to the dynamic information age, learning style, and so on with the learning process for the change. Learning style, and other information is the first time users log on the system when the registration information to fill in the content, such as knowledge of the presentation preference, and so on.

C. The module about Student's cognitive state

Students can study the performance of their level of awareness of the most obvious link is the practice test, this paper links for the students reasoning. According to the theory of cognitive knowledge and understanding of students at grade 6, 1-6 by the increasing degree of order, but here is not from 1 to 6 levels of difficulty, we say that each has a different level of difficulty, and so on the difficult The title should do with the difficulty of the questions on the analysis of the small students to understand the different levels of contribution; In addition, the students in question do, so the speed title reflects his familiarity with the knowledge that if a student in a long time To make a title, we can not think he has a good grasp of this knowledge. Through the above analysis, apart from the question hierarchical distinction, in recognition inference extent, the students also consider the following two aspects: the difficulty of examination questions and students doing the test of time.

For the different level of test, we use five grade, they are difficult, more difficult, medium,more easily, easily, for 0.25,0.5,1,2,4.

For students to do the test of time, we have provided for the two time periods: the normal time and the longest time. Students in the usual time to do on the title, we said he was familiar with this knowledge; normal and he was among the longest title to show he was not very familiar with; If more than most of its time that we are not familiar with this knowledge. As the students a familiarity with the knowledge itself is ambiguous, so we use the membership function to express familiarity. The membership function to use the fuzzy focus of the most commonly used functions under the S function.The define of the S function is in Figure 2:

$$S(x;\alpha,\beta,\gamma)=\begin{cases}0 & \text{当 } x\le\alpha\\ 2\left(\dfrac{x-\alpha}{\gamma-\alpha}\right)^2 & \text{当 } \alpha\le x\le\beta\\ 1-2\left(\dfrac{x-\gamma}{\gamma-\alpha}\right)^2 & \text{当 } \beta\le x\le\gamma\\ 1 & \text{当 } x\ge\gamma\end{cases}$$

Figure 2: The define of the S function

D. The module about Student's Emotional state

Students estimate is mainly used to monitor the emotional students in the process of emotional changes, appropriate adjustments on the part of students learning a great impact, but also the most difficult to deal with. This is because the psychological situation very difficult for students from certain acts reflected, even face-to-face teaching is not easy to see, let alone through an intermediary for the machine to analyze. For these reasons, the mood in this article for students to estimate the only student in the state of mind in practice. The definition of a problem to do with the maximum allowed time for the T, students do a title that used to t, students answer with Answer, said no students in a row prompted the wrong number for n.

Rule 1: If t ≤ T and Answer = True Then No prompt and give the students praise End;

Rule 2: If t ≤ T and Answer = False Then
If n <5 Then No prompt and comfort the students
Else prompte and encourage the students
End
End;

Rule 3: If T <t ≤ 2T and Answer = True Then No prompt and encourage and praise the students End;

Rule 4: If T <t ≤ 2T and Answer = False Then prompte and encourage and comfort the students End;

Rule 5: If t> 2T Then prompte and encourage and comfort the students End.

The content of prompt are two aspects of the students of the wrong reasons and the right method.

E. The module about Student's study procedure

Students process information module is used to record the progress of students, students used to describe the domain knowledge. To address this issue this paper, the concept of a map of the methods were used to map the concept of the students said the knowledge and expertise, the concept plan for each node of the knowledge that this course, students, of course, expert knowledge is knowledge of the concept of a map of the sub-plans, Students in the learning process, the concept of middle school students knowledge of the plans will continue to expand and eventually form a framework in line with the requirements of the concept plan. In this paper, the students knowledge on the concept map of each node to give students an awareness that the level of fuzzy sets, so that the fuzzy set by the judge to identify students at the cognitive level of knowledge.

The steps are as follows: First of all, when students first visit of a course of a knowledge-point, in its concept map plus a point, the point was marked Show, said the students have read, but the system does not know whether he Have learned the content, and would also like to record their way of learning media, the node is an isolated point. Second, students do exercises to strengthen the knowledge of the point of understanding that a revised level of awareness of fuzzy sets, this time looking for expert knowledge in the concept plan at this knowledge nodes, and then in accordance with expert knowledge in this concept map node And the other nodes to connect the relationship between the nodes connected to the students knowledge of the existence of the concept of map-related node, the node is not an isolated point in the. This marks the node Show changed Learned.

Through the nodes of the level of awareness of the concept of fuzzy sets and said that graph of the relationship between knowledge point, the system will not only be able to understand the students knowledge on the point of learning, but also has a diagnostic capability to help students find possible reasons for the error, so Better learning.

F. The module about Student's study interest

Interest in learning module is the analysis of the interest, a design student of the best teaching strategies. Its interest in two areas: the presentation of knowledge and knowledge of education. Students interested in the estimates used in this article at the same time three ways for students interested in estimation. 1.'s Inquiry showed that the user preferences; 2. By tracking the user's behavior; 3. The students through the use of certain knowledge of the presentation and educational way to learn after the effects of good or bad.

IV. REASONING METHOD IN THE SYSTEM

Students through conduct of the information obtained is a personality of its uncertain reasoning, Bayesian, fuzzy logic is one of the more effective methods of reasoning, students are often used in the model. In this paper, a fuzzy logic.Student interaction with the system more suitable for use fuzzy sets to that than the probability of a more rational manner. Because the learning is a continuous process of change and transformation, the system only through the student movement to carry out similar reasoning, and fuzzy logic reasoning for this is more appropriate.Bayesian networks need to know the conditions of probability and the probability of thing before, which is difficult to obtain the probability of the need to carry out a study of Bayesian network before reasoning. This process was time-consuming and costly, and it is very difficult, very difficult to practice. The fuzzy logic approach provides a direct and less complex reasoning, should be easier to achieve.With the result of the introduction of fuzzy theory and Bayesian network is not accurate, but a probability very close to its value.Using fuzzy sets and fuzzy logic that the students complete the model can be replaced after Bayesian network or other methods, and do not need major changes.

The uncertainty of fuzzy logic theory is a major concern and quantitative reasoning, these quantitative reasoning and the use of natural language and contains many of the ambiguous meaning of the word. Fuzzy logic is fuzzy set theory, the basic developed. The tradition of an object that is a feature of the collection is in the form of function, sometimes referred to as identification function.

If an object is a collection of elements, it features the function of 1; if not a collection of elements, while its value is 0 characteristic function, the definition can be summarized into the following characteristics of the function: $\mu_A(x): x \in \{0,1\}$ .This binary logic of the problem is that we live in a mathematical simulation and not the world. In the real world, things are usually one or the other is not, therefore, gave rise to the concept of fuzzy sets:

In the domain of the so-called X on a fuzzy subset A refers to any $x \in X$ has a corresponding number $\mu_A(x) \in [0,1]$, and called X belonging to the fuzzy subset A membership, which means that the map

$$\mu_A: X \in [0,1] \qquad x \in \mu_A(x)$$

A map μ_A, also known as the membership function, fuzzy sets are also often referred to as fuzzy sets. That's usually the following method for $X = (x_1, x_2, \ldots x_n)$, then A fuzzy set of X can be expressed as:

a) $A=$

$$\sum_{i=1}^{n} \frac{\mu_A(x_i)}{x_i} = \frac{\mu_A(x_1)}{x_1} + \frac{\mu_A(x_2)}{x_2} + \ldots + \frac{\mu_A(x_n)}{x_n}$$

V. CONCLUSION

It can achieve individuational education with a cognitive level in traditional network intelligent tutoring system, but it can not be perceived emotion of students in the process of teaching, it can not be achieved emotional aspects of individuational education. The main topics of this item is the key of the application to the cognitive theory and emotional theory, we update the building method about the student's learning model and teaching

strategy in intelligent navigational learning system in order to be a great intelligent network teaching platform with cognitive and emotional interaction.

REFERENCES

[1]Guoliang Yang, Zhiliang Wang. Emotional modeling research. Automation technology and application,2004.

[2] Kaicheng Yang ,Student model and learning activities designed,2006.

[3]http://www.etr.com.cn/

[4]http://211.152.9.125/

[5]http://disted.tamu.edu/edtclink.htm

[6]http://carbon.cudenver.edu/~lsherry/pubs/issues.htm

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 088-093

Model Based Security Policy Assessment for E-Business Environment

Wang Chu, and Yanli Feng
Shandong Institute of Business and Technology, Yantai, China
Email: sdchuw @ 163.com

***Abstract*—The key to profitability for e-business is ensuring data integrity, service availability, and user information confidentiality along the entire e-services chain. Both staffs and IT system components need to compare secure policy with performance in an e-business environment. Currently, most efforts set focus on e-business process analysis and value-chain analysis, little attention is put on the secure policy compliance assessment. This paper presens a model based security policy assessment approach that integrates fault tree analysis technology and top-down architecture driven system analysis method. The assessment process includes security attribute scenarios generation, e-business security model construction, fault tree based threat model construction, and security policy evaluation. It can be used to analyze the security policy for the e-business environment from two different perspectives: 1) Compliance analysis between security policy and e-business security model, intended to elicit all possible discrepancies; 2) Adequacy analysis of security policy for identified threats, aiming at verifying and demonstrating whether the security policy are appropriate for the emerging secure risks.**

***Index Terms*—E-business, Security policy assessment, Architecture driven system analysis, Fault tree analysis.**

I. INTRODUCTION

Currently e-business systems are evolving in socio-technical systems, where human and organization factors assume a more and more critical role in the secure operation of the systems. A socio-technical system is represented as a complex network of interrelationships between human and technical systems that includes hardware, software, actors, data, and rules. The key to profitability for e-business is ensuring data integrity, guaranteeing service availability and protecting confidentiality along the entire e-services chain. Security of e-business system has become increasingly important. In general, the security goals for e-business environment will be to ensure that the sensitive information are only used for the purpose for which they have been distributed, to protect them against theft, duplication etc. Organizations that adhere to a strict secure policy management and compliance do more than improve corporate financial performance. They also reduce the chance of legal exposures and liabilities due to negligent protection of key corporate assets[1].

Security is a property of the e-business system that prevents undesired outcomes. The effectiveness of security mechanisms can only be judged in the context of the overall system. A security policy with regard to control over access and dissemination of information must be precisely defined [2]. Security policy can range from organizational level rules and practices that regulate how the sensitive data are managed, protected, and distributed within an e-business organization to detailed sets of actions or practices regulating the processing of sensitive information and the use of resources by the hardware and software of an e-business system.

Both staffs and IT system components need to compare secure policy with performance in an e-business environment. Compliance assessment can identify potential security risks and also highlight aspects of security policy that may have become outdated or impractical. These regularly policy analysis ensure tight alignment between security policy and the e-business environments.

Policy compliance begins by identifying what information is available online, and which part of that data rates a specific level of protection. All of the system components, such as business components, service components, networks, databases, and so on, are the basis for creating a security policy – the blueprint by which an organization decides the level of protection for any online resource.

Currently, most efforts set focus on e-business process analysis and value-chain analysis, little attention is put on the secure policy compliance analysis. Research existing on security compliance focuses on testing application protocols rather than the whole socio-technical system [3]. The most promising method is goal-based refinement, since it provides clarity about the nature of goals and an abstract model to manage the analysis process. However, it is difficult to apply this approach directly to the e-business environment, because it do not allow the expression of system topology and e-service deployment. Both these are security critical features of an e-business system[2].

Security policy assessment is about discovering risk in what are not specified, what are specified inadequately, or what are outdated. In this paper, we propose a model based security policy assessment approach that combines fault tree analysis technology with top-down architecture driven system analysis method to assess the security policy for e-business environment. Its purpose is to probe undocumented secure actions or practices for identified secure risks. The assessment process includes: security attribute scenarios generation, e-business security model construction, fault tree based threat model construction, and security policy evaluation.

The contributions of this paper consist in the following aspects: 1) Fault tree analysis technolory is used to construct the threat model at component level; 2) Utility tree technology is used to construct the security attribute scenarios; 3) Architecture driven system analyzing method is used to generate the e-business security model. In addition, we define the relationships among the models formally that can be used to develop automated assessment tool.

This paper is organized as follows. Section II outlines the related work on security requirement and risk analysis approach. Section III details the model based security policy assessment process. Section IV outlines a case study. Finally, section V rounds up the paper with a conclusion and ongoing work.

II. Related Work

Howard Chivers reviews development methods for secure systems in the light of security requirements found in emerging distributed systems, these requirements are particularly concerned to protect sensitive information such as workflow records, licenses or electronic products. There are two aspects: security goals and available security design practice. Howard Chivers also discusses the security concerns (Confidentiality, Integrity, Availability), information system policies, and risk management procee model[2].

Hao Chen and Jean-Pierre Corriveau discuss the current security testing categories and standards, as well as common security testing approaches. They propose an original scheme to design a compliance testing system for the security of online banking. Their proposal aims at suggesting to testers how to design security testing and identify potential vulnerabilities in current online banking systems. They summarize criteria and ideas that may be useful for the creation of a compliance testing approach for the security of online banking systems. Such an approach should be designed according to the security standards and policies of banks[3].

Gyrd Brædeland and Ketil Støen advocate asset-oriented risk analysis as a means to help defend user trust. Their paper focuses on a net-bank scenario, and addresses the issue of analysing trust from the perspective of the bank. The proposed approach defines user trust as an asset and makes use of asset-oriented risk analysis to identify treats, vulnerabilities and unwanted incidents that may reduce user trust [4].

N. Nayak et al. introduce the concept of core business architecture for a service-oriented enterprise. The business architecture of a service-oriented enterprise is adequately represented through five main architectural domains: business value, structure, behavior, policy, and performance. Business rules describe ‘‘what’’ is required, rather than ‘‘how’’ it should be implemented. In many cases, a single business rule must be implemented in multiple aspects of an implementation. Business rules may extend various aspects of other kinds of models, such as: operation models, specification models, business performance models, information models [5].

III. Model Based Security Policy Assessment

Policy compliance and vulnerability management are important functions of an e-business organization. Security compliance may include analysis and testing of the system for conformance to a set of security policies. It is important that a security evaluation of e-business environment be performed using official standards [1].

A. Introduction to Architecture Driven System Modeling and Analyzing Method

Architecture defines a set of components, specifies a topological pattern for their interconnection, and imposes constraints on them, which captures the vital parts of a structure in an organized manner and a practical means for managing a complex system, such as software system or a business. Hence, architecture is introduced into e-business modeling and analyzing. The architecture driven system modeling and analyzing method follows the traditional “divide-and-conquer” method of defining architecture that consists of three activities: goal decomposing, architecture defining, and validating [6].

• Goal decomposing: The objective of this activity is to divide the system goal into a number of sub-goals and assign them to components. In this activity, “Responsibility-Assignment” relationships between the system goal and the components are created, and they are called γ-relationships.

• Architecture defining: The objective of this activity is to construct architecture to achieve the system goal. Determining choreography of the components is the major work of this activity. “Take-Part-In” relationships between the components and the constructed architecture are created, and they are called β-relationships.

• Validating: The objective of this activity is to check whether the constructed architecture meets the system goal or requirements. In this activity, “Achieved-By” relationship between the system goal and the constructed architecture is created, and it is called λ-relationship.

An e-business system model constructed by means of the architecture driven system modeling approach can be written as a pair $e\text{-}M=(B_s, R)$, where, B_s is a set of components containing business objectives (i.e., business components, service components, actors); R is a set of relationahips between components.

To model and assess security policy, we need to analyze: a) e-business objectives and their realizations (e-services and IT components), b) interdependencies among assets, and c) the risks that threats business objectives. Generally, security policy and business objectives are too abstract to be analyzed directly. They should be decomposed into operational policies and business components step by step according to the e-business architectures at different abstract levels.

We use architecture driven system analyzing method to assess the security policy for e-business environment. The assessment process starts from the business objectives and secure goals. We refine them by iterative decompositions. Iteration is needed so that the operational e-service components and actors are identified.

The fulfillment of business objectives might be disrupted by the occurrence of uncertain negative events (i.e., threats, un/intentional events, incidents) direct or indirectly (by disrupting the supporting assets).

B. Security Policy Model for E-business Application

There are multiple definitions possible for e-business secure policies as follows[7]:

1) Policy helps to define what is considered valuable, and specifies what steps should be taken to safeguard those assets.

2) Policy is defined as the set of laws, rules, practices, norms, and fashions that regulate how an organization manages, protects, and distributes sensitive information, and that regulates how an organization protects system services.

3) Access to a system may be granted only if the appropriate clearances are presented. Policy defines the clearance levels that are needed by system subjects to access objects.

4) In an access control model, policy specifies the access rules for an access control framework.

Firewalls, encryption servers, card keys, VPNs and similar technologies do not eliminate risk so much as they shift it from one part of the e-business system to another. Poorly chosen passwords, card keys and misconfigured network devices easily foil access control and authentication. Encryption only protects data while in transit. It is still at risk before transmission and after it is arrived. Even worse, the encryption stream can be disrupted, corrupting traffic and causing expensive data integrity repairs [1].

Security is a whole-system consideration and not just confined to software systems, network infrastructure, or staffs. Security policy assessment cannot be considered in isolation. Security policy for an e-business environment should be planned and analyzed systematicaly, which involves all of the business components.

We define security policy model for an e-business environment as follow in order to analyze the consistency between security policy and the e-business environment.

Definition 1. A security policy model is a 3-tuple $e\text{-}P=(G_s, A, M)$, where:

- G_s is a set of secure goals at different abstract levels;
- A is a set of actions or practices in order to achieve secure goals;
- M: $G_s \rightarrow \mathbb{P}(A)$, mapping secure goals into actions or practices.

C. Fault Tree Based Security Policy Assessment

The security policy assessment process consists of following steps:

- Security attribute scenarios generation;
- E-business security model construction;.
- Fault tree based threat model construction;
- Security policy evaluation.

(1) Security attribute scenarios generation.

Security goals are often summarized as 'Confidentiality, Integrity, Availability'[2]:

- Confidentiality means that information embodied in either data items or executable programs is accessible only by authorized parties.
- Integrity have a range of possible interpretations: maintain consistency, prevent inappropriate modification, detect modification or allow recovery.
- E-business strongly depends on the availability of e-services, an organization should be able to ensure the continuity of its business objectives.

The first step in the security attribute scenarios generation involves identifying key security attributes or concerns associated with e-business environment. Identifying key security attributes serves to limit the number of hazards that can be considered at one time and to place the hazards in e-business context. To help in the process of identifying security attributes, a set of global security goals for e-business system are used as a starting point. Global security goals are derived from general organisational concerns on security.

Similar to [8], we use utility tree to derive concrete security attribute scenarios that are used to generate fault tree. Utility tree provide a top-down mechanism for directly translating the e-business secure goals into concrete quality attribute scenarios. Before we can assess the securrity policy for an e-business environment, these secure goals must be made more concrete. The utility tree contains utility as the root node. The quality attributes of performance, modifiability, security, and availability are the high-level nodes immediately under utility. In this paper, security factors are refined to specific sub-factors that are concrete enough for security analysis. The utility tree guides the analysts to check whether the e-business environment satisfies the security scenarios at the leaves of the utility tree.

The security goal is refined to three sub-goals: Information confidentiality, Data integrity, and Resource availability, and in turn, such goals are refined into more concrete goals. For example, information confidentiality may be refined to two concrete scenarios: Customer database authorization works 99.99% of time and Credit card transactions are secure 99.99% of time. The security attribute scenarios are used to identify threats to an e-business environment and the security policy of an e-business organization.

(2) E-business security model construction.

Security policy assessment for e-business environment should meet the following prerequisites:

- Understand the IT infrastructure and the e-business objectives it support. Some system components and information resources are more valuable than others, and not all of it needs to be protected equally.
- Understand component-level solutions such as firewalls, authentication and encryption with integration technology that maximizes effectiveness for the whole e-business environment.

We need to create a detailed security model based on e-business system architecture specification to show how specific measures meet the secure goals, such as using up-to-date anti-virus, anti-spyware software for actor platform.

In paper [9], we propose an architecture centric development approach to e-business application that leads to e-business plan and e-service model. We use these models at different abstract levels to define the e-business security model.

Definition 2. Given e-business system model $e\text{-}M=(B_s, R)$, its security model is a 4-tuple $e\text{-}S=(B_s, R, C, I)$, where:

- B_s is a set of components containing business objectives (i.e., business components, service components, actors);
- R is a set of relationahips between components;
- C is a set of secure measures;
- I: $B_s \rightarrow \mathbb{P}(C)$, mapping business objective into secure measure(s) that satisfies related secure goals.

(3) Fault tree based threat model construction.

Ivan Victor Krsul discuss the action features of vulnerability and the consequence features in [7]. The action features of vulnerability include: 1) The exploitation of the vulnerability can result in a user observing sensitive data in violation of expected policy; 2) The exploitation of the vulnerability can result in a user destroying data in violation of expected policy; 3) The exploitation of the vulnerability can result in a user modifying data in violation of expected policy; 4) The exploitation of the vulnerability can result in a user creating data in violation of expected policy.

The consequence features are: 1) The exploitation of the vulnerability can result in a change of availability of the system; 2) The exploitation of the vulnerability can result in the disclosure of information in violation of expected policy; 3) The exploitation of the vulnerability can result in the misrepresentation of information; 4) The exploitation of the vulnerability can result in repudiation of information; 5) The exploitation of the vulnerability can result in a change of integrity of the system; 6) The exploitation of the vulnerability can result in the loss of confidentiality of information.

For each identified security attribute scenario, we use fault-tree to construct threat model in order to discover possible vulnerabilities in an e-business environment and in security policy. This process largely relies on the judgement and experience of the analysts involved in the assessment process.

Fault-tree analysis is concerned with discovering the e-business system states which are potentially hazardous. For each identified hazard (condition with potential for causing an accident), a fault-tree is produced which traces back to all possible situations which might cause that hazard. Fault-tree analysis can be applied at different abstraction levels from a business objective through to an analysis of a software component. Fault-trees include and/or operators which allow hazardous conditions to be combined. A fault tree model precisely documents which failure scenarios have been considered and which have not. Fault tree analysis can be used to support engineering and management decisions, trade-off analysis and risk assessment. A vulnerability is a weakness with respect to an asset or group of assets which can be exploited by one or more threats. The weaknesses range from easily guessed passwords to misconfigured or unauthorized devices/information resource, to improper user activity. There are six types of vulnerabilities that can exist in any system, and these are: Physical, Natural, Hardware/Software, Media, Communication, and Human [10].

Given a secure goal (security attribute scenario) "g", "$\neg g$"expresses a failure that is used to construct a context specific fault tree. The fault tree diagram is used to identify the unusual, but possible combinations of component failures. This method is (also called flaw hypothesis) based on propagation of errors due to vulnerabilities in order to find possible ways to affect the e-business system. The idea is to hypothesize possible flaws, and then check whether these hypotheses are true.

For each threat consideration (flaw hypothesis), we identify a list of cutsets. A cutset is a set of basic events. In this paper, the threat model is constructed based on the cutsets and the component diagrams. The threat model is defined as following:

Definition 3. Given e-business system model $e\text{-}M=(B_s, R)$, the threat model is a 3-tuple $e\text{-}V=(B_s, V, K)$, where:

- B_s is a set of business objectives (i.e., business components, service components, actors);
- V is a set of threats;
- K: $B_s \rightarrow \mathbb{P}$ (V), mapping component into the related threats.

(4) Security policy evaluation.

For e-business systems, the security policy evaluation is concerned with discovering the weaknesses or risks on e-business operation. Once we have constructed the security policy model, the e-business security model, and the threat model, we can assess the security policy for the e-business environment from two different perspectives:

– Compliance analysis between security policy and e-business security model, intended to elicit all possible discrepancies.

– Adequacy analysis of security policy for the identified threats, aiming at verifying and demonstrating whether the security policy are appropriate for the identified threats. In essence, the adequacy analysis may be viewed as a proactive rather than reactive approach to security so that the e-business organization can mitigate risk and threats.

Due to the more and more emerging secure threats, once a security policy is in place, it is necessary to update it at regular intervals and to assess hosts, networks, applications and databases for vulnerabilities systematically and regularly.

Definition 4. (Compliance). Given e-business security model $e\text{-}S=(B_s, R, C, I)$ and security policy model $e\text{-}P=(G_s, A, M)$, for every b ($b \in B_s$), $\exists g$ ($g \in G_s$), and g is the secure goal of b, if $I(b) \subseteq M(g)$ then we say that $e\text{-}P$ is implemented by $e\text{-}S$ properly, that is, $e\text{-}S$ complies with $e\text{-}P$ properly, denoted as $e\text{-}S \prec e\text{-}P$.

Definition 4 means that if every secure measure in e-business environment complies with organization's

security policy then the security policies are implemented by the e-business environment properly.

Definition 5. (Security Policy Adequacy). Given security policy model $e\text{-}P=(G_s, A, M)$ and threat model $e\text{-}V=(B_s, V, K)$, for every b ($b\in B_s$), $\exists g$ ($g\in G_s$), and g is the secure goal of b, if for $K(b)$ there exist according security action(s) in $M(g)$ then we say $e\text{-}P$ is adequate for $e\text{-}V$, denoted as $e\text{-}P \hookrightarrow e\text{-}V$.

Otherwise, if $\exists b$ ($b\in B_s$), $\exists g$ ($g\in G_s$), and g is the secure goal of b, if for $K(b)$ there does not exist according security action(s) in $M(g)$ then we say $e\text{-}P$ is not adequate for $e\text{-}V$.

Likewise, we can give the definition to e-business security model suffice as following:

Definition 6. (E-business Security Model Suffice). Given e-business security model $e\text{-}S=(B_s, R, C, I)$ and threat model $e\text{-}V=(B_s, V, K)$, for every $b(b\in B_s)$, if for $K(b)$ there exist countermeasures in $I(b)$ then we say $e\text{-}S$ is sufficient for $e\text{-}V$, denoted as $e\text{-}S \twoheadrightarrow e\text{-}V$.

We can use fault tree and business scenarios to assess e-business security environment.

Theorem Given security policy model $e\text{-}P=(G_s, A, M)$, e-business security model $e\text{-}S=(B_s, R, C, I)$, and threat model $e\text{-}V=(B_s, V, K)$, if $e\text{-}S \prec e\text{-}P$ and $e\text{-}P \hookrightarrow e\text{-}V$ then $e\text{-}S \twoheadrightarrow e\text{-}V$.

According to Definition 4, Definition 5, and Definition 6, the theorem is hold obviously.

The theorem means that if the security policy has containted sufficient actions/practices to the identified threats and the e-business environment has been verified to comply with the security policy, then the e-business environment is secure.

The security policy assessment leads to an assessment report that contain two parts: 1) Discrepancies between the security policy and the e-business environment; 2) Inadequacy of the security policy to the emerging secure risks. Once inadequacies in the security policy and e-business environment are discovered, the e-business organization should revise the security policy and select appropriate security technologies for e-business environment, the latter must be tested to demonstrate they meet the security demands of the relevant level(s).

IV. CASE STUDY

In an e-business environment, disclosure of the customer data to outsiders is an example of a security related unwanted incident during a normal transaction. The e-business environment focuses on the security aspects of two separate functionalities of the actor platform: 1) The actor authentication mechanism; 2) The secure payment mechanism, which is based on the use of digital certificates and communication channel encryption.

The use of name/password for actor identification can have undesirable consequences if a malicious actor captures another actor's ID (name/password). This actor is able to embarrass the legitimate user exploiting the captured ID by logging into the platform. In this case study, we will discuss the security policy assessment. Suppose that the secure goal is "protecting confidentiality of the actor's ID during the entire business transaction process". The e-business security model, security policy model, and threat model are illustrated in Figure 1.

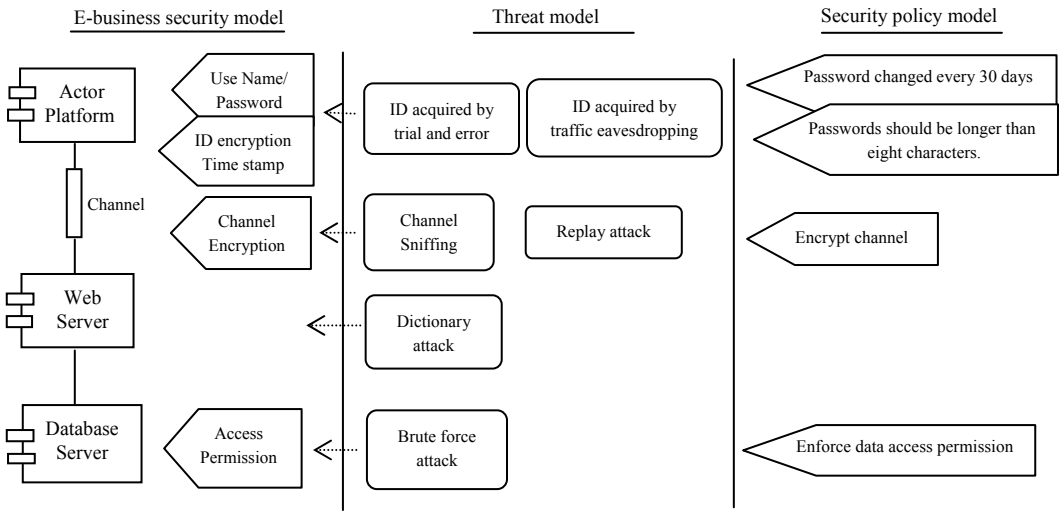


Fig. 1 Security policy assessment for e-business environment

A legitimate actor logs into the e-business system with the name/password, during business transaction, user's name/password may be leaked, Fig.1 shows the possible vulnerabilities and the current securerity policy. Due to the limited space, the construction steps of the models are omitted, we only give the result.

In Fig.1, the security policies are implemented by the e-business environment very well, but the the security

policy against the emerging secure risks are not adequate. There are not proper actions against the dictionary attack in the security policy and in the e-business environment. Hence the security policy should be revised and new secure thnology should be selected for the e-business environment, such as account timeout or auto-lock, and so on.

V. Conclusion and Future Work

Security is a property of the e-business system that prevents undesired outcomes. Both staffs and IT system components need to compare secure policy with performance in an e-business environment. Compliance assessment can identify potential security risks and also highlights aspects of security policy that may have become outdated or impractical. In this paper, we propose a model based security policy assessment approach that combines fault tree analysis technology with architecture driven system analysis method to assess the security policy for e-business environment. Compared with current asessment methods, our approach has following characteristics: 1) Fault tree analysis technolory is used to construct the threat model at component level; 2) Utility tree technology is used to construct the security attribute scenarios; 3) Architecture driven system analyzing method is used to generate the e-business security model. In addition, we define the relationships among the models formally that can be used to develop automated assessment tool.

The future work includes developing automated security policy assessment tool and risk management tool

References

[1] "Secure E-business". Technical Report, Internet Security Systems, 2000.

[2] Howard Chivers. "Security and Systems Engineering". Technical Report, University of York, Heslington,. York, 2004.

[3] Hao Chen, Jean-Pierre Corriveau. "Security Testing and Compliance for Online Banking in Real-World". Proceedings of the International MultiConference of Engineers and Computer Scientists, 2009.

[4] Gyrd Brædeland and Ketil Støen. "Using Risk Analysis to Assess User Trust– A Net-Bank Scenario". C.D. Jensen et al. (Eds.): iTrust 2004, LNCS 2995, 2004. pp. 146–160.

[5] N. Nayak,M. Linehan, A. Nigam, et al. "Core business architecture for a service-oriented enterprise". IBM Systems Journal, Vol 46, No 4, 2007. pp.723-742.

[6] Wang Chu, Depei Qian. "Semantics Based Enterprise Modeling for Automated Service Discovery and Service Composition". *Proceedings of 2007 IEEE Asia-Pacific Services Computing Conference*, Tsukuba, Japan, December 11-14, 2007. pp.439-445.

[7] Ivan Victor Krsul. "Software Vulnerability Analysis". PhD Dissertation, Purdue University, 1998.

[8] Daniela Barreiro Claro, and Patrick Albers, and Jin-Kao Hao. "Web services composition". *In Semantic Web Service, Processes and Application*. Springer, 2006. pp.195-225.

[9] Wang Chu, Depei Qian. "A Component-Oriented Development Approach to E-Business Applications". *Proceedings of IEEE International Conference on e-Business Engineering.* October 22-24, 2008, Xi'an China. IEEE Press. pp.45-52.

[10] Stilianos Vidalis and Andy Jones. "Using Vulnerability Trees for Decision Making in Threat Assessment". Technical Report CS-03-2, University of Glamorgan, June 2003

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 094-097

Application Research of Embedded Web Technology in Traffic Monitoring System

Rui Li [1], and XiangQiang Xiao [2]
[1] Department of Information Engineering, AnHui Communications Vocational & Technical College, Hefei, China
liruilary@gmail.com
[2] School of Machinery and Automobile Engineering, Hefei University of Technology, Hefei, China
bluesprint1978@yahoo.com.cn

***Abstract*—the paper mainly discusses design of hardware & software for embedded web server with Heterogeneous network seamless connectivity and implement of key technology. It contains transplantation of embedded Linux operating system, design of embedded web server, transplantation of database and implementing method of main functions. Remote monitoring is realized to traffic information collection, monitoring traffic conditions, traffic control, information published and communication of traffic data by using combining EWS technology with Internet. The results indicate that the intelligent traffic control technology based on embedded web technology can achieve the integration of a wide range of information collection and it breaks through the traditional traffic monitoring technology for designing traffic monitoring system and provides the advanced technology based on embedded Web for designing modern traffic monitoring system. The testing showed the traffic monitoring system based on embedded Web technology has the good real-time and high reliability and good scalability and anti-interference performance. Meanwhile, it has also laid a good foundation for the further study of a new type of intelligent traffic information collection system.**

***Index Terms*—embedded web technology, traffic monitoring system, EWS system, information integration, intelligent transportation**

I. INTRODUCTION

With rapid economic development in China, transportation has increasingly become an extremely important component in the national economy and daily life. So it is very essential to build a modern intelligent traffic control system in order to resolve the traffic congestion of roads and reduce accidents. And video monitoring and traffic information transmission in this system plays an important role. Intelligent traffic for solving urban traffic management has become the people's consensus such as advanced and sophisticated video surveillance system as an important component of intelligent transportation for image acquisition, on-site snapshot, after taking of evidence and other important tasks. Monitoring systems are usually installed on the expressway, traffic junctions, toll stations and other key places according to the actual needs of current traffic monitoring. All the information is integrated to the monitoring center.

The project is supported by Transportation department of Anhui

At present, the traffic management monitoring systems are based on the IPC as a host computer, and deploy dedicated monitoring configuration software. This method is not only costly, inefficient, but also troublesome for the system to update, and specialized training for management personnel, and restricted by space-time and geography. Moreover, some information cannot be shared for public information services.

With the rise of the Internet technology, embedded Web technology goes into the mainstream at present, and CGI script and Web server support the program running on an embedded device. The managers can manage and monitor situations of traffic through the Web browsers.

This paper presents a method that combines embedded WEB technology with Internet to implement remote traffic monitoring through Web Server applications solidified in embedded ARM processor. Therefore managerial personnel can have the remote real-time monitoring of traffic management through Web browsers without time and geographical constraints as shown in Figure 1. Time-consuming effort of traditional local monitoring, as well as deficiencies in equipment maintenance, is effectively overcome and efficiency of traffic management is greatly improved. In order to classify information, embedded Web Server applications consist of different treatments of classification according to the types of information that are then properly displayed in the browsers. Attention is given especially for the effective separation of the confidential and public information [1,2].

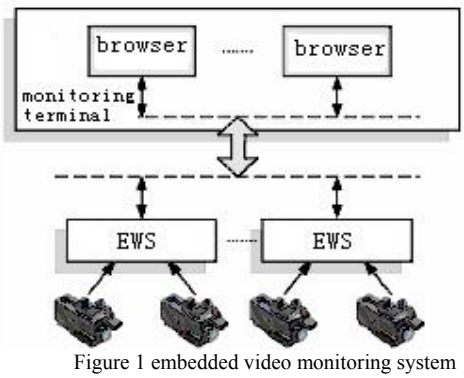


Figure 1 embedded video monitoring system

II. SYSTEM DESIGN

EWS system is composed of embedded web server hardware & software system and traffic monitoring system, as shown in Figure 2. The hardware mainly

consists of two main components: an embedded Web server and bus controller. The bus controller is for region of traffic monitoring to design, and each control point is identified corresponding to the embedded WEB server (EWS), and the hardware structure and software system of EWS are determined according to the corresponding control volume[3,4].

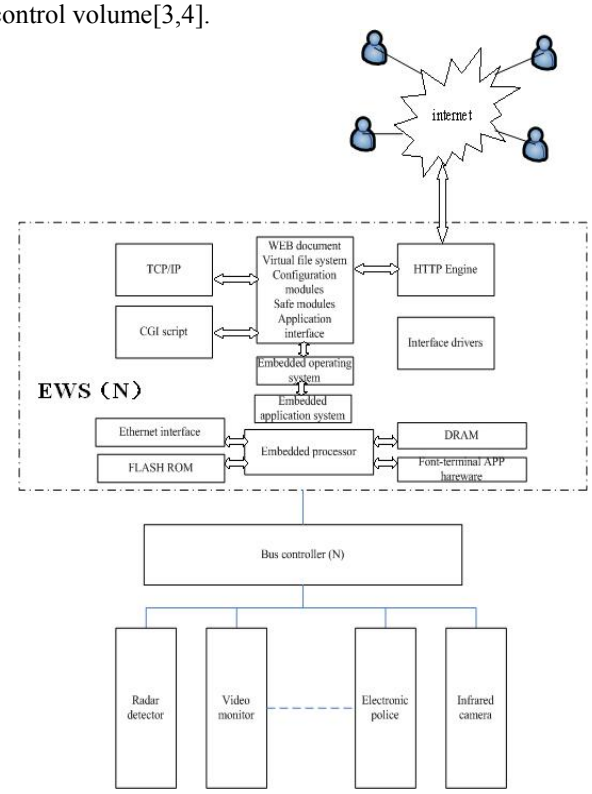


Figure 2 composition of system

EWS Hardware System

EWS hardware system includes an embedded ARM processor, FLASH, ROM, DRAM, Ethernet port, front-terminal application system components and bus controller, as shown in Figure 3.The processor is Samsung S3C44B0X ARM7 that is suitable for quite poor working environment, supports uCLinux operating system, runs faster than 8-bit and 16-bit processor[11].

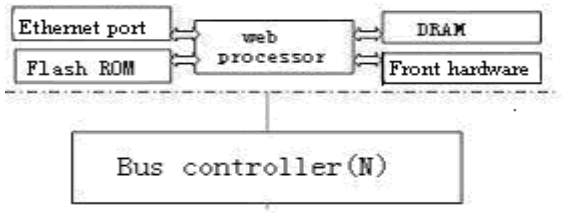


Figure 3 EWS hardware system

EWS Software System

EWS software system includes HTTP engine, TCP / IP protocol, CGI script, virtual file system, configuration module, security module, application interface module, embedded operating system, embedded application, interface driver, embedded database SQLite[14].

CGI(Common Gateway Interface) design is the most important in EWS software system and defines the interface standard between Web server and CGI script. CGI programmed by C language is embedded with Html script. While CGI is executed, some special ports can be operated and the results are displayed in the browsers. The special operation is as follows: send some information to a Web server from the client, and put the received information into environment variables, then go to start the specified CGI script in order to complete specific task, and CGI script obtains the relevant information from the environment variable and decode it to some serial port, dispatch commands concerned to monitoring module on-site, then return the output with HTML format to browsers through Web server[10,13]. Its workflow is as shown in Figure 4.

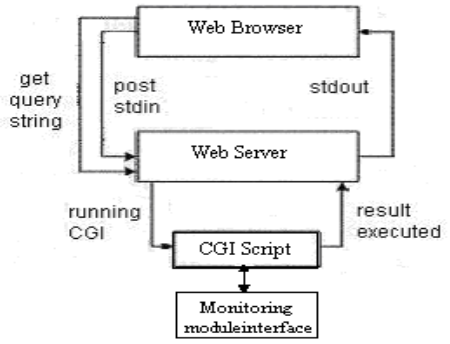


Figure 4 CGI workflow

Transplanting Design of EWS Operating System

The selection of the operating system is essential in the EWS design. uClinux operating system is chosen according to the actual needs of the system as well as its stability and reliability.

uClinux is a branch of Linux--- Micro-Control-Linux. It is mainly aimed at non-MMU processor and supports multi-task with a complete TCP / IP protocol stacks as well as multiple network protocols. uClinux also supports a variety of file systems.

First, Linux 2.4.20 is chosen for trimming & transplanting its kernel and devices drivers, configuring uCLinux and processor, patching the compiler as well as building the cross-compiler environment. The kernel trimmed is mainly a set of configurations for hardware platform, file system, network protocols and so on. Moreover, retains serial console and common tools & commands. Then Web Server, Telnet daemon and other popular applications are transplanted. Usually the kernel is compressed into FLASH, and extract it into RAM to run with file system with romfs, and executable file format with flat as well as run-time library uCLibc with simplified. Typical drivers include the console terminal, serial devices, and block device drivers with file system. Module technology dynamically loaded in uCLinux is used for drivers in developing and debugging ---makes drivers compiled into the kernel, boots directly to load, and also supports uCLibc DLL[2,13]. As shown Figure 5.

Applications
Linux Kernel
Boot loader
Hardware Device

Figure 5 EWS software system

SQLite Database Design

There are a large number of equipments in traffic Monitoring System at fields and some monitoring data should be stored in a database. Embedded database SQLite can be more easily used in embedded systems although SQLite is a lightweight relational database. It also supports multiple tables and indexes, transactions, views, triggers and a series of user interfaces and drivers. SQLite database is accessed By calling the C language API interface(Mainly call three API functions).

Monitoring System

Monitoring system includes infrared cameras, electronic police, video surveillance devices, Radar detectors, bus controller and etc. The traffic information from the monitoring equipments is transmitted to the EWS through the bus controller. Then displayed in browsers through the EWS to implement traffic management and remote real-time monitoring[6].

III.CONFIGURATIONS OF EWS HARDWARE

EWS uses a Samsung S3C44B0X processor, K9F1208U0M Flash with 64M, HY57V561620SDRAM with 32M, CS8900A Ethernet port equipped with RS232 and USB1.1 standard, multiple 4--wire serial port with extended to connect multiple peripheral devices, I / O ports with expanding[13]. As shown in Figure 6.

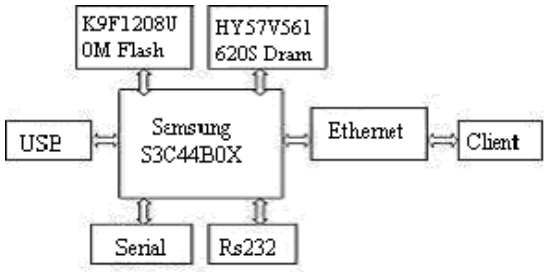


Figure 6 hardware configurations

IV. DETAILED DESIGN OF MONITORING SYSTEM

The decentralized monitoring and centralized management are generally used in traffic monitoring system. The entire system contains a monitoring center, a number of scattered remote monitoring terminals and telecom media.

The remote monitoring terminals include the video transmission terminals and the data transmission terminals[7].The video transmission terminal is generally a infrared camera or monitor and the video signals from it are compressed by the MJPE or JPEG through the multi-function bus controller to form data frames that can be transmitted in the network, then the data frames are transferred into the EWS through the internal bus. The data transmission Terminals are such as the electronic police, radar detectors, vehicle detectors, variable speed limit signs, variable information panels and etc. These devices are responsible for data acquisition, data storage, and data transferring through the multi-function bus controller, and then transfer it to the EWS through the internal bus. Therefore managers are able to have the real-time monitoring of remote traffic management through EWS connected to the Internet / Intranet[5].

Software of the monitoring system consists mainly of information acquisition module and device control module. The information acquisition module is mainly to complete information collection on-site, and also to implement status information collection from its own peripheral devices. The information collected is uploaded to the host monitoring and displayed with the query. Thus the current field conditions, as well as equipment working status, are taken into consideration for appropriate control measures such as alarm. The device control module is mainly for controlling equipment operating parameters and information publication. The operating parameters mainly refer to the sampling period, sampling rate, alarm thresholds, mode of equipment work and etc[8,9]. The specific process control is shown in the main program of the monitoring system in Figure 7.

As the EWS server starts running, the main module program is executed to access between the EWS and browsers, and the EWS sends data concerned to the sub-module program by using POST method. When a client

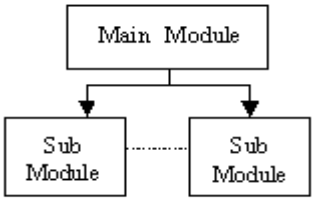


Figure 7 the main program of the monitoring system

from a browser sends the requests to the EWS, the corresponding sub-process is booted by the server daemon, then the sub-process meets each specific request. The sub-process is module mainly composed of several processing modules: ① general function module. ② static text processing module. ③ CGI module and error processing module[12]. Results from the sub-process module are returned to the client as the HTTP response message after The EWS application completes the corresponding operation.

V. SYSTEM FUNCTION AND TESTING

The embedded Web traffic monitoring system is based on embedded Web technology as the core. Meanwhile, it is combined with traffic information acquisition, traffic surveillance, traffic control, information publication and other traffic control functions. Then the traffic data are collected, stored, managed, transmitted, analyzed, and displayed. The traffic managers or decision makers are provided with these data for decision making and management on the traffic situation. For example, when a vehicle runs the red light, the electronic police system detects the vehicle through the ground induction coil; its detector is triggered, at the same time the signal controller issues a "red light" signal to the electronic police system. When both conditions are matching, an image of the vehicle, with the relevant monitoring

information is taken as illegal driving. For another example, while the radar detectors transmit the radar beam in the direction of the road and receive the reflected echoes of the vehicle, the speed of the vehicle is measured by echo analysis. If the speed exceeds the setting, it will direct the camera to obtain the relevant speeding monitoring information. With heterogeneous network seamlessly connected with embedded gateway design as well as the realization of key technologies, these traffic monitoring information are integrated to the EWS through a serial port, parallel port, USB and so on. On-site monitoring traffic information is sent to the web browsers via a EWS. So wherever the monitoring personnel are, the target region can be monitored and managed as long as it is connected to the Internet network. At the same time public travel information services can also be easily provided. Diagram of system function is as shown in Figure 8. The traffic information web browser is as shown in Figure 9.

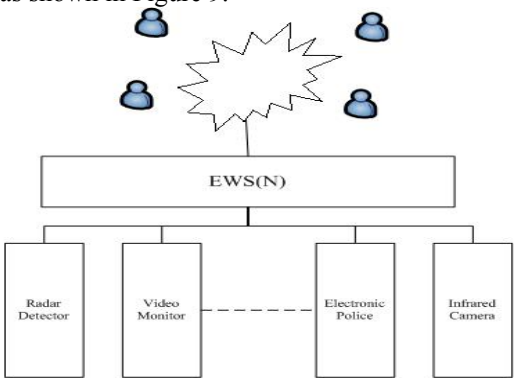


Figure8 diagram of system function

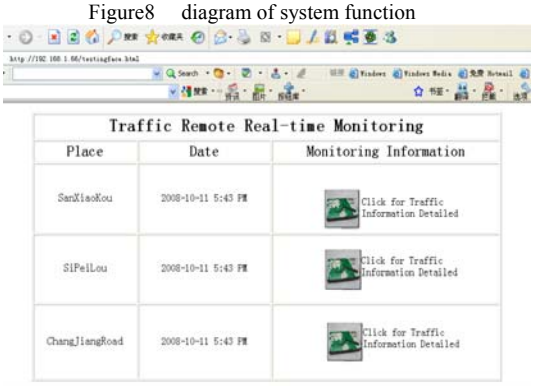


Figure 9 UI of traffic information web browser

VI. CONCLUSION

The system adapts an embedded Web server technology to implement data collection and video monitoring through using modular structure and heterogeneous network seamlessly connected. Therefore the managers can perform the traffic management and monitoring without the geographical and environmental limitation and the public information services are also provided. Intelligent traffic control system based on the embedded Web technology is widely used along with the development of the intelligent traffic.

The traffic monitoring system based on embedded Web technology possesses the low power consumption, high integration, real-time efficiency, software solid-state, and easy scalability. Moreover, it is able to effectively manage the increasing complexity of system resources, and makes some hardware virtualization. As data exchange with the outside interfaces only through the data transfer protocol, so the system is provided with the good device independence, replacement ability, commonality, and scalability. It easily makes the system flexibly to be assembled, replaced, and upgraded. Therefore significant savings in investment and an increase of benefits are for the traffic monitoring system based on embedded Web technology.

REFERENCES

[1] Yuan Yi , “A remote video surveillance system based on embedded Web server technology” [J]. Power System Technology, 2000,24(5):71-73.

[2] Huang BuY, Zheng AnPing, Liu GuoMei , “Web technology implement based onμCLinux” [J]. Electronic Design & Application,2003,12:87-90.

[3] Coelho C N,da Silva D C,Padrao W C, “Reengineering embedded systems for the internet”[C]. 15th Triennial World Congress, Barcelona ,Spain, 2002 IFAC:761-772.

[4] Wan JiaFu,ZhangWenFei,Zhang ZhanSong, “Principles and applications of network monitoring system” [M].China Machine Press,2003:178-289.

[5] Fu BaoChuan, Ban JianMin, “Design of remote monitoring system based on Web embedded ” [J]. Control & Automation(Embedded and SOC),2005,21,7-2:58-60.

[6] Li JuGuang,Zhen Gen, Jiang ZeMing, “Embedded linux system development in details———based on EP93XX ARM” [M]. Tsinghua University Press,2006:157-256.

[7] Mao Yong, Jin WenZhen, “Remote fault diagnosis system based on Web server” [J]. Application of ElectronicTechnique,2003,(3):56-59.

[8] Huang Ying, Xiao Xu,Wen JiBo,“Design of remote monitoring system based on embedded Linux” [J]. Electronics Engineer, 2002,28(4) :11-13.

[9] Zong Chang-Fu, Yang Xiao, Wang Chang, et al, “Driving intentions identification and behaviors prediction in car lane change” [J]. Journal of Jilin University (Engineering and Technology Edition), 2009, 39: 27-32.

[10] Farid M.N, Kopf M, Bubb H, et al, “Methods to develop a driver observation system used in an active safety system” [J].VDI Berichte,2006, 1960: 639-650.

[11] Han XiaoTao,Yin XiangGen, Zhang Zhe, LI Wei, ” Review of embedded web server technology and its application in power system” [J] Power System Technology 2003,(5): 58-62

[12] JIA LiNa, Wang Zhen,” Application of Embedded Web Server in Intelligent Residential District” [J] Instrumentation Technology 2004,(5): 23-24.

[13] LI Yong,” Application and realization of CGI in embedded WEB server”[J].Microcomputer Information, 2008,(30):110- 111.

[14] Li ShuiYang,Han XiaoTao, ”Application of embedded WEB server technology” [J]. Journal of higher correspondence education(natural sciences) 2003, (6)Vol. 16 No. 3: 47-50.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 098-101

Rapid Detection of Heavy Metal Contents in Fruits by Laser Induced Breakdown Spectroscopy

Mingyin Yao[1] , Muhua Liu[1*] , Lin Huang[2] , and Jinhui Zhao[1]
[1]College of Engineering, [2]College of Bioscience and Engineering
Jiangxi Agricultural University
Nanchang, Jiangxi, 330045, China
E-mail:mingyin800@126.com, suikelmh@shou.com, huanglin213@126.com, zjhxiaocao@sohu.com,
leizejian123@163.com,liqiulian620@163.com

***Abstract*—To detect heavy metal contents in fruits rapidly, the Citrus Nanfeng tangerines pericarp and flesh have been analyzed by laser induced breakdown spectroscopy. Line emissions from five heavy metal elements, Pb, Cd, Hg,Cr and As, have been clearly extracted. Their intensities correspond to relative concentrations of these elements contained in the analyzed samples. The results demonstrated that the species and contents of heavy metal in fruits can be identified by their LIBS spectra, and the heavy metal contents in the inner of fruits are more than the outer. This analysis showed efficient discrimination between heavy metals from different parts of a single fruit by laser induced breakdown spectroscopy spectra.**

***Index Terms*—fruits, heavy metal contents, rapid detection, LIBS**

I. INTRODUCTION

In recent years heavy metals elements contained in fertilizer and environment contamination have been used in agriculture widely. Some deleterious heavy metals elements, such as plumbum (Pb), cadmium (Cd), hydrargyrum (Hg), chromium (Cr) and arsenic (As), are transmitted into fruits and other farm produces. Some of them are transited into high toxic compound going with food in body. Most of them have the characteristic of accumulation and longer half life, which brings acute or chronic toxicity reaction so that teratogenicity, cancer-causing and mutagenicity come into being [1, 2]. So, the detection and identification of heavy metals elements in foods, such as fruits, has been an important problem around the world.

The traditional means of detecting the heavy metal contents in fruits has been used by inductively coupled plasma atomic emission spectrometry (ICP-AES), flame atomic absorption spectrometry (FAAS), and so on[3, 4]. But the samples need to be digested by concentrated acid in the above ways. The pre-treatment processing is complicated and time-consuming, the real-time detection can not be implemented, and the secondary contamination appears easily.

Laser-induced breakdown spectroscopy (LIBS) is basically an emission spectroscopy technique where atoms and ions are primarily formed in their excited states as a result of interaction between a tightly focused pulse-laser beam and the material sample [5]. One of the important features of this technique is that it does not require any sample preparation, unlike conventional spectroscopic analytical techniques. Samples in the form of solids, liquids, gels, gases, plasmas and biological materials [6-10] can be studied with almost equal ease. LIBS has rapidly developed into a major analytical technology with the capability of detecting all chemical elements in a sample, of real- time response, and of close-contact or stand-off analysis of targets.

In this paper, the characteristic spectrum of elements Pb, Cd, Hg,Cr and As in orange pericarp and flesh were collected by laser-induced breakdown spectroscopy. Based on our results we propose an innovative strategy to detect heavy metal contents in fruits rapidly.

II. MATERIALS AND METHODS

A. Experimental samples

Fifty Citrus Nanfeng tangerines from Jiangxi province were chosen as experimental objectives. Each orange was divided into pericarp and flesh at the same dimensions. That is, the total number of samples is 100. The pericarp and flesh have 50 samples respectively.

B. Experimental Setup

Experimental setup sees Fig.1.The Nd:YAG (yttrium-aluminum-garnet) nanosecond laser (BeamTech, Nimma-200) provides pulses of 10 ns duration at 1064 nm and a repetition rate of 10 HZ. The laser beam is focused on the surface of a sample using a single lens of 200 mm focal length. Typical maximum pulse energies at the ablation surface were 200 mJ/pulse. To have a fresh spot for each laser shot, the sample is fixed on a move platform. Optical emission from the LIBS plasma was collected by a 2 m steel encased multimode fiber (core diameter=400 um) .This fiber was coupled to an eight-channel AvaSpec spectrometer equipped with a 2048 pixel CCD (charge coupled device) detector (AVANTES B.V., 2048-USB2-RM) which provided complete spectra coverage from 200 to 1100 nm with a resolution of 0.07 nm at 315 to 417nm. The AvaSpec spectrometer was controlled by personal computer (PC)

Project Supported by Natural Science Foundation of China (NO.30972052)
*Corresponding author: Muhua LIU,Email:mingyin800@126.com, suikelmh@shou.com

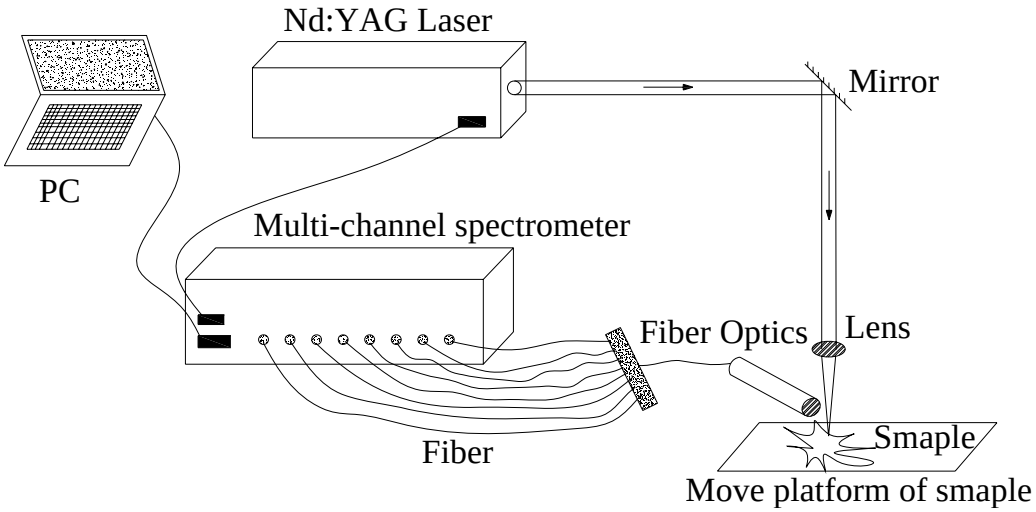


Figure 1. The experimental setup of LIBS

running manufacturer-provided software. LIBS spectra were acquired at a delay time of 1.28 *us* after the ablation pulse. For each pericarp or flesh sample 20 spectra were taken. Each individual spectrum was accumulated over 100 laser shots. Spectra were recorded continuously in order to keep the experimental conditions as identical as possible for all samples.

For each individual spectrum, the line intensity was extracted for each of elements Pb, Cd, Hg,Cr and As. For given experimental conditions, we can consider the line intensity of an element is proportional to the relative concentration of the element. The average of 20 spectra of individual sample was extracted as the line intensity of the element. Line intensities of elements in sample provide a profile of relative concentrations of the elements.

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

Fig.2 and Fig.3 show a typical LIBS spectrum from Citrus Nanfeng tangerines pericarp and flesh respectively. Five elements have been included in our consideration: Pb, Cd, Hg,Cr and As. From each individual spectrum, line intensity was extracted for each of the five elements. Look first at Fig.2, the spectrum is dominated by emission from Cr and Pb. To a lesser extent, emission from Cd, As and Hg is visible. For the same experimental conditions, continuous emission is much more obvious in the pericarp spectrum than the flesh spectrum. That can be clearly seen by comparing Fig.2 and Fig.3.

The 238.12 nm As line, 247.77 nm Cr line, 274.85 nm Cd line, 280.20 nm Pb line and 302.15 nm Hg line were chosen as the characteristic spectra. Each line was collected for 20 spectra, and each spectrum was accumulated for 100 laser shots. For the data processing, I_{ij}^{k} is used to represent a specific raw line intensity, where k represents different samples(k=1….,100), j refers to an individual spectrum(j=1….,20), and i refers to a specific element(i=1….,5).In order to present a relative concentration of the element. The mean value $\bar{I}_{i}^{k}$ of line intensities of an element are calculated for 20 spectra of a given sample.

$$\bar{I}_{i}^{k} = \frac{1}{N}\sum_{j=1}^{N} I_{ij}^{k}$$

The average intensity of emission line is given in Table 1.

In Table 1, for the same element, the mean values of LIBS relative line intensity is more in pericarp than in fresh. That indicates the relative contents of elements are more in pericarp than in fresh. This result fits well with the property of heavy metals in fruits, in which heavy metals accumulate in outer of the fruits more than the inner. The contents of heavy metals are less in inner of the fruits. So, line intensities associated to the elements for a given fruit samples provide a profile of relative concentrations of the elements. Further studies are underway to detect the heavy metal contents from outer to inner in fruits by LIBS spectra.

TABLE I.
MEAN VALUES OF LIBS RELATIVE LINE INTENSITY OF ELEMENTS

Samples	Relative line intensity (dimensionless)				
Nanfeng tangerines	As238.12nm	Cr247.77nm	Cd274.85nm	Pb280.20nm	Hg302.15nm
Pericarp	35.222	220.391	40.522	242.915	20.935
Flesh	3.491	64.705	5.413	46.731	5.123

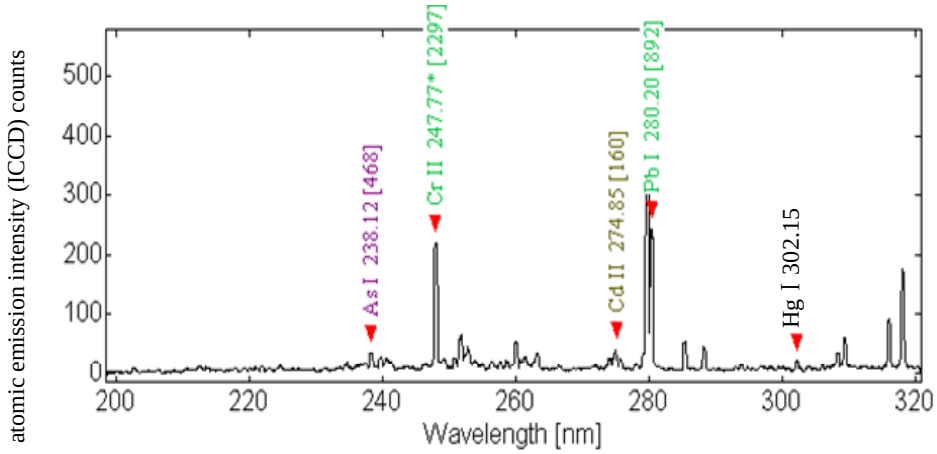


Figure 2. LIBS spectrum of Citrus Nanfeng tangerines pericarp with the Pb, Cd, Hg,Cr and As emissions identified

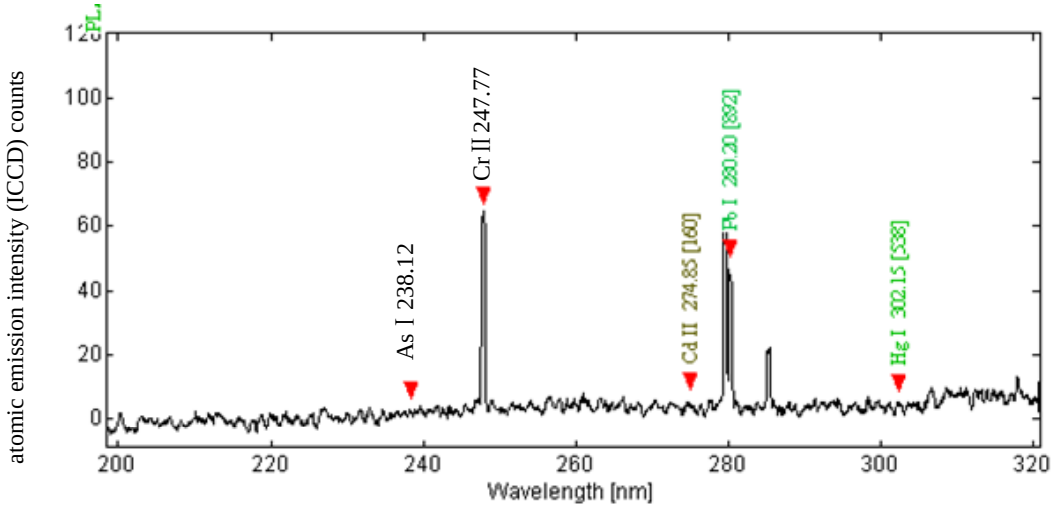


Figure 3. LIBS spectrum of Citrus Nanfeng tangerines flesh with the Pb, Cd, Hg,Cr and As emissions identified

Ⅳ. CONCLUSIONS

In this paper, we have extracted the characteristic spectra of Pb, Cd, Hg,Cr and As in Citrus Nanfeng tangerines pericarp and flesh by LIBS. The results demonstrated that the species and contents of heavy metal in fruits can be identified by their LIBS spectra, and the heavy metal contents in the inner of fruits are more than the outer. The profile of relative concentrations of the elements can be associated with their LIBS spectra. Such relative concentrations can be precisely determined by LIBS. However, much progress needs to be achieved before LIBS can be considered as a practical application. Experiments are currently underway in our laboratory utilizing several different samples of fruits to address these problems.

ACKNOWLEDGMENT

This work was supported by a grant from National Natural Science Foundation of China.

REFERENCES

[1] Pence N S,Larsen P B,Ebbs S D,etal, “The molecular physiology of heavy metal transport in the Zn/Cd hyper accumulator,”*Proc Natl Acad Sci*, vol.97,pp.4596-4560,2000.

[2] Alam MGM.,Snow ET.,Tanaka A.,*Arsenic and heavy metal contamination of vegetables grown in Samta village, Bangladesh* , The Science of the Total Environment, 2003, 308:1-3.

[3] SONG Ji-li,CHEN Xu-wei,YU Zuo-wen,et a1,*Determination of harmful compound fertilizers by metal in organic-inorganic ICP-OES method*,Chemical Fertilizer Industry，2006，33(6)27-28．(in Chinese).

[4] HUANG Dong-gen.LIAO Shi-jun，ZHANG Xin-quan，et a1, *Analysis of six elements in rice field soilby ICP-MS*,Enviromental Monitoring in China， 2005， 21(3)：31-34． (in Chinese).

[5] Sun Q, Tran M, Smith B W, et al, *Zinc analysis in human skin by laser induced-breakdown spectroscopy*,Talanta, 2000, 52 (2):293-300.

[6] Michela C, Gabriele C, Montserrat H, et al, *Application of laser-induced breakdown spectroscopy technique to hair tissue mineral analysis*,App. Opt, 2003, 42 (30): 6133-6137.
[7] Kumar A, Sharma P C,*Uses of LIBS technology in biological media*, Proc SPIE, 2006, 6377: 11-17.
[8] Matthieu B, Laurent G, Jin Y, et al, *Spectral signature of native CN bonds for bacterium detection and identification using femtosecond laser-induced breakdown spectroscopy*, Applied Physics Letters, 2006, 88(6): 063901 (1-3).
[9] Matthieu B, Laurent G, Jin Y, et al, *Femtosecond time resolved laser-induced breakdown spectroscopy for detection and identification of bacteria: A comparison to the nanosecond regime*,Journal of Applied Physics, 2006, 99 (8): 084701 (1-9).
[10] Vincent Juvé, Richard Portelli, Myriam Boueri, et al, *Space-resolved analysis of trace elements in fresh vegetables using ultraviolet nanosecond laser-induced breakdown spectroscopy*, Spectrochimica Acta Part B: Atomic Spectroscopy, 2008, 63(10): 1047-105
[11]

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 102-105

A Kind of Low Complexity LDPC Decoder

Hang Jiang[1], Chun Xu[2], Qin Zhong[3], and Guifeng Zhong[3]
[1] School of Electronic Science and Technology, Huazhong University of Science and Technology, Wuhan, China
Email: liqing0922@163.com
[2] No.1 Middle School Attached to Central China Normal University, Wuhan, China
[3] School of Information Engineering, Wuhan University of Technology, Wuhan, China

***Abstract*—A kind of (1008, 3, 6) rules LDPC decoder was designed in this paper, in order to solving the high complexity of LDPC decoder hardware implementation, which adopted Min-Sum decoding algorithm and part-parallel structure. The minimum module in the traditional check function unit (CFU) was improved, which could reduce the complexity of the hardware realization. In the end, the LDPC decoder achieved the desired effect through the comparison of performance analysis, which laid a good foundation for the practical application.**

***Index Terms*—LDPC, Min-Sum, part-parallel, minimum module, low complexity**

I. INTRODUCTION

In modern communication system, error-correcting code is an important way of improving the channel transmission reliability and power utilization. While low-density parity-check codes (LDPC code) is a kind of error-correcting code which gets most close to Shannon limit [1]. However, LDPC codes makes the most of the advantage of its superior performance when the code length is long, which means that the scale of encoding and decoding circuitry will also be proportionate increased relatively and chip area also will be larger. Nowadays, the study of the current LDPC decoder hardware implementation focuses mainly on maintaining the decoding performance while reducing the complexity of the LDPC decoder in the FPGA or DSP devices. Regarding to the problem that LDPC decoder occupies too much hardware resources, this paper puts forward a new minimum module, with the hardware resources greatly reduced without any loss in decoding speed and performance. This method was verified and had some practical value to the LDPC decoder.

II. LDPC DECODING ALGORITHM

In the hardware implementation, Sum-Product Algorithm and Min-Sum Algorithm are widely used in LDPC decoding algorithm. Min-Sum Algorithm is a simplified form of Sum-Product Algorithm. The following example is based on the channel of additive white Gaussian noise (AWGN) and binary phase shift keying (BPSK) modulation. The updating and iteration formula of bit nodes and check nodes can be expressed as [2]:

$$L(b_{ij}) = L(b_j) + \sum_{i' \in C(j)\backslash i} L(c_{i'j}) \quad (1)$$

$$L(c_{ij}) = \prod_{j' \in B(i)\backslash j} \alpha_{ij} \bullet \min_{j' \in B(i)\backslash j} \beta_{ij'} \quad (2)$$

Where $L(b_{ij})$ is the log-likelihood ratio information of bit node; $L(c_{ij})$ is the log-likelihood ratio information of check node $\alpha_{ij} = \mathrm{sign}[L(b_{ij})]$, $\beta_{ij} = |L(b_{ij})|$; $L(b_j) = 2c_j / \sigma^2$ is the initial information of the channel.

So the whole decoding process only includes addition and executing minimum value. Take the (N, K) rules of LDPC code as an example, row is d_c and column is d_b. The complexity of Min-Sum Algorithm is shown in Table 1.

TABLE I. ANALYSIS OF THE COMPLEXITY OF MIN-SUM ALGORITHM

Algorithm	number of operation (add, subtract)	number of Comparison	number of look-up table
Min-Sum Algorithm	$2 d_b N$	$2 d_c K$	0

III. DESIGN OF THE OVERALL STRUCTURE OF DECODER

Because the serial structure has only one degree of parallelism and its latency is too large, the general LDPC decoder rarely takes the serial structure. Parallel structure increases the degree of parallelism so that coding rate is very desirable, but led to a corresponding exponential increase in hardware resources [3] [4].Considering the resource consumption of hardware implementation and processing speed, compromising the advantages and disadvantages of parallel and serial structure, a kind of part-parallel structure decoder is designed in this paper, which not only improves the speed but also doesn't over-occupied chip resources.

A. Part Of The Parallel Structure

The overall structure of the decoder is shown in Fig.1. Its FPGA structure consists of six modules: control input data module, operation control module, check nodes update module, bit nodes update module, data storage module and matrix calibration module. First we convert the data whose length is 3bits coming from the data channel into 1008bits as a data and partition into parallel

Hang Jiang: (1989-10).

24 way. Each path contains $1008/24=42$ bits. So $w_b=24$. Bit Function Unit (BFU) is needed to carry out the calculations of Update module. Each BFU will address 42bits data in each iteration calculation. The number of Check Function Unit (CFU) is half the BFU ($w_c=12$). And each CFU will also deal with 42bits in one iteration calculation.

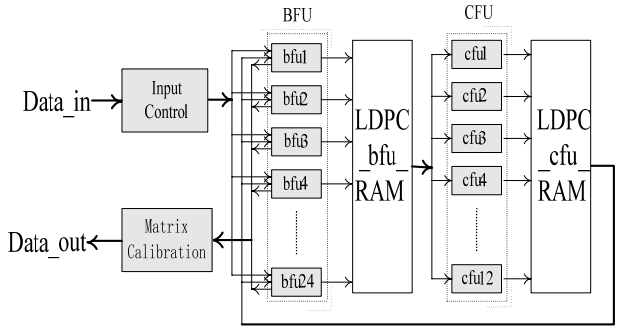


Figure 1. The overall structure of Decoder

B. Bit Function Unit(BFU)

According to formula 1, BFU is to complete the job of summation operations. We take six-channel data as a large group, so that is divided into 24 / 6 = 4 large sets of data. Thereby, it realizes part of parallel structure and avoids the huge resource consumption of addressing the 24-channel data at the same time [4]. The 6-chan-nel data c_1,c_2,c_3,c_4,c_5,c_6 from check node update module are processed by adder and overflow process, and then go through a look-up table (LUT). The new data of BFU is b_1,b_2,b_3,b_4,b_5,b_6 . The chart of BFU structure is shown in Fig.2.

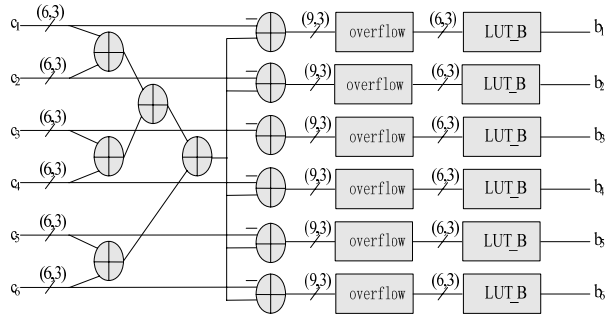


Figure 2. BFU structure

C. Check Function Unit (CFU)

According to formula 2, we can see, each CFU needs to complete the tasks as follows: 1) seek the minimum absolute value of each data; 2) seek the product of the value of data symbol; 3) merge the minimum and the value of symbol into a new data. CFU structure is shown in Fig. 3.

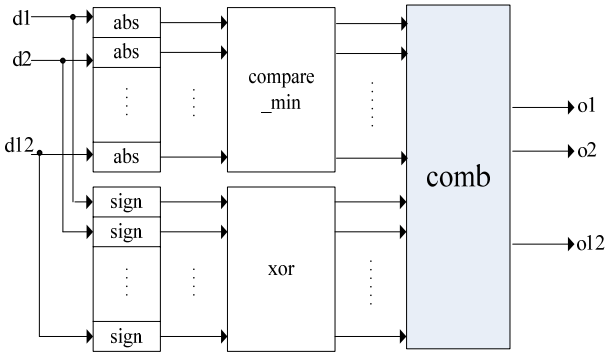


Figure 3. CFU structure

The compare_min Module carries out the operation of computing the minimum absolute value of each data in formula 2. Its structure is shown in Fig.4. The xor module is to complete the product of the value of data symbol involved in computing the value of each data, which is equivalent to xor operation, as shown in Fig.5.

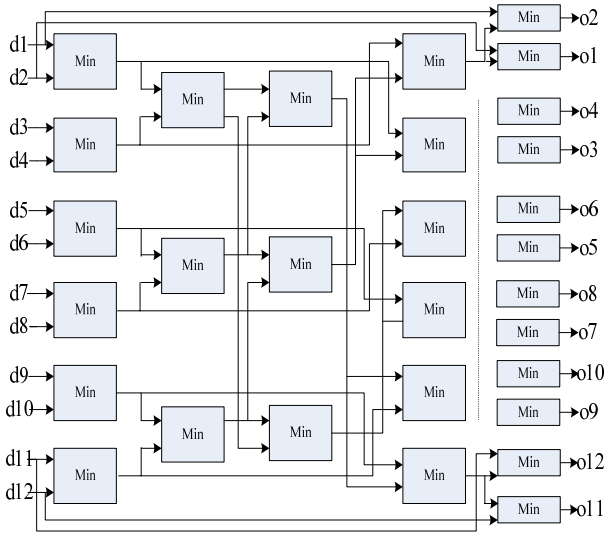


Figure 4.Compare_min Module

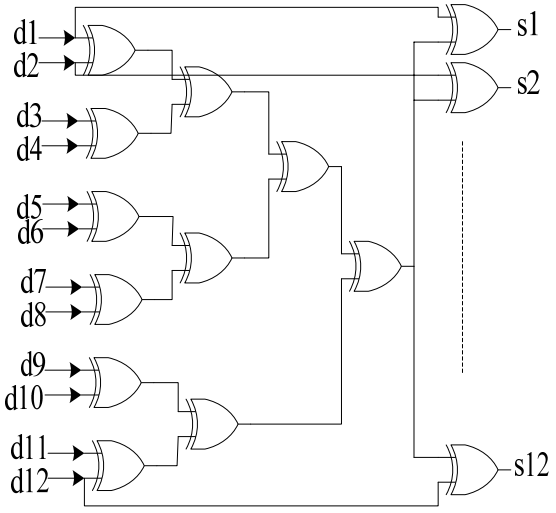


Figure 5. xor module

D. Improved CFU

As can be seen from Fig.4, there are 12 node values participating in the calculation. When the NO.1 node (o1) is the minimum, we need to put NO. 2～12 node value into the formula to calculate; when the NO.2 node (o2) is the minimum, we need to put NO.1, NO.3～12 node value into the formula to calculate, and so on, when the NO.12 node (o12) is the minimum, we need to put NO.1～11 node value into the formula to calculate. With this method the number of comparators used is 30. Obviously, when the CFU is larger, the scale of this structure is also doubled.

In fact, the function of the minimum module can be understood as follows: there is always a minimum and a sub-minimum value in all the relevant data that check node received. When the minimum is from the first k-bit node, then the first k-bit check node is to receive the sub-minimum value of relevant nodes and the remaining check nodes is to receive the first k-bit node value. Therefore, the compare_min module simply completes the operations of computing the minimum and the sub-minimum value. The structures of seeking the minimum and the sub-minimum value adopt four kinds of basic structure, as is shown in Fig.6.

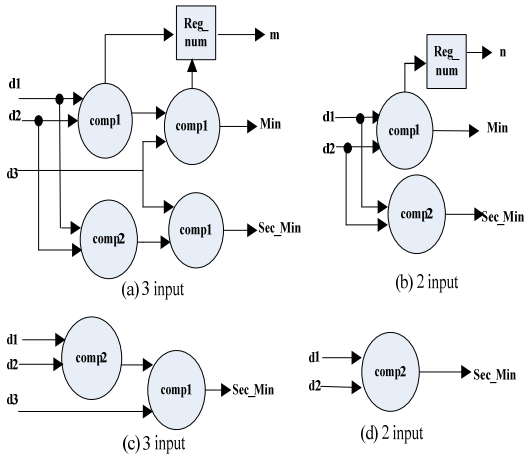


Figure 6. The basic structures of seeking the minimum and the sub-minimum value

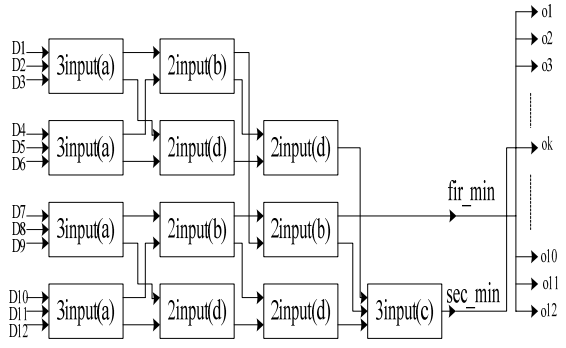


Figure 7. The improved compare_min module

Fig.6 (a) is the 3-input and 2-output structure, which outputs the minimum and sub-minimum value and records the input data m corresponding to the minimum value; Fig.6(b)is for the 2-input and 2-output structure, which outputs the minimum and sub-minimum value and records the input data n corresponding to the minimum value; Fig.6(c) and Fig.6(d) output the minimum for3-input and 2-input data respectively. Applying the structures of seeking the minimum and the sub-minimum value to the compare_min module, the improved compare_min module is shown in Fig.7.

Reg_num will update the input data 1 corresponding to the minimum according to the last record and the next (a) (b) structure. Then for a 12 data input, l is a 4bits record. Assuming the input data is k corresponding to the minimum, the improved compare_min module requires 4×4+3×2+1×2+4×1=28comparators, which is reduced 6.7%. When the check node degree is larger, the advantage of such resources will be more obvious. It is shown in TABLE 2.

TABLE II. THE COMPARISON OF THE NUMBER OF COMPARATORS

Number of Input nodes	Comparator of Traditional minimum module	Comparator of Improved minimum module	Reduced Comparator
12	30	28	（28-30）/30=6.7%
20	90	62	（90-62）/60=31.1%
26	180	118	（180-118）/180=34.4%

From Fig.4 and Fig.7 we can see that the traditional compare_min module uses six cycles, and the improved compare_min modules uses (2 +1 +1 +2 =) 6 cycles too. That is, these two structures are the same in decoding rate.

IV. SIMULATION AND ANALYSIS

A. Decoder Verification on FPGA

The (1008, 3, 6) Rules of the LDPC decoder is programmed with Verilog HDL language, achieving the traditional CFU structure and the improved CFU structure of LDPC decoder respectively. We select Altera-Statix Ⅱ series of FPGA, EP2SGX60EF1152C3 and compile them in the Quartus5.1. The hardware resource utilization is shown in table 3 [5].

TABLE III. HARDWARE RESOURCE UTILIZATION

resource	utilization volume		Utilization factor	
	traditional CFU	*improved CFU*	*traditional CFU*	*improved CFU*
ALUTs	18,574	16,392	38.4%	33.7%
Total RAM bits	46,300	47,352	1.82%	1.86%

From Table 3 we can see that a small increase in the number of register in exchange for a greatly reduced system logic resources. Thus we validate the effectiveness of the designed CFU structure in this paper.

In addition, the two decoders of different structures obtain the maximum clock frequency of 82.7MHz and

82.3MHz respectively. In the circumstances of allowing error, we also validate the correctness of the two kinds of LDPC decoder in decoding rate.

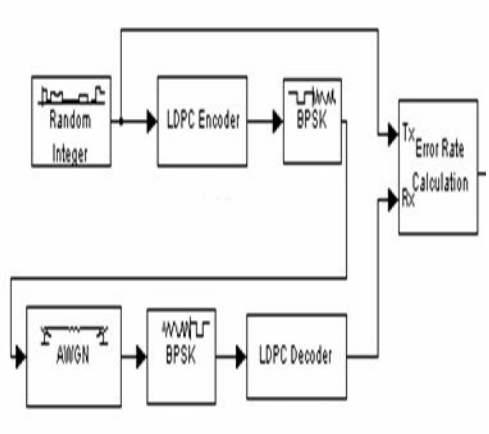


Figure 8. Simulation platform of LDPC decoder

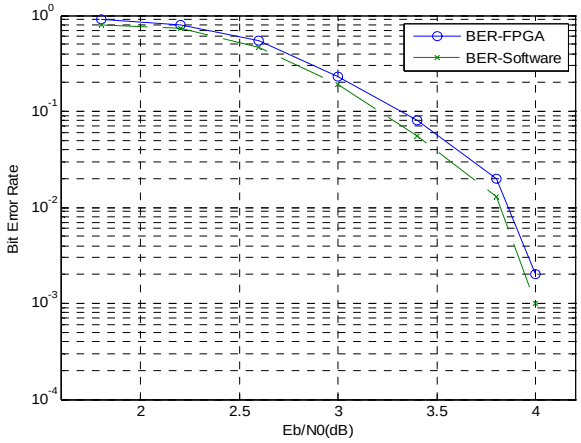


Figure 9. Comparison of FPGA simulation and Software simulation

B. Performance Comparison

Software simulation platform of LDPC decoder designed in this paper is shown in Fig.8. 1008bits binary message sequences from random source are sent to the channel encoder-LDPC encoder. The encoded information is modulated digitally, then entering the AWGN channel. Signal at the receiving terminal are demodulated and LDPC decoded, then becoming the output signal. To test the performance of LDPC codes, we obtain the BER using bit-error-rate statistics module through comparing code word from decoder and source.

Fig.9 is the comparison of the actual test result and software simulation result of decoder. A six quantitative decoder is used because there will be some noise in hardware testing. On the whole, the actual test performance is in line with expectations and the throughput of decoder is up to 142Mbps.

V. CONCLUSION

The new kind of LDPC decoder designed in this paper adopts a simple and efficient Min-Sum decoding algorithm and part-parallel structure, and improves the minimum module, which reduces the utilization of hardware resources significantly. Through simulation analysis, under the premise of guaranteed decoding speed and performance, the consumed logic resources are reduced significantly, which has a certain practical value indeed.

REFERENCES

[1] Bin Zhang. ”Research on B3G key technology —— non-regular LDPC decoder of part parallel structure design and implementation” [D]. [Master’s degree thesis].Chengdu: University of Electronic Science and Technology .2006

[2] Kazunori Shimizu, Tatsuyuki Ishikawa, Nozomu Togawa, Takeshi Ikenaga and Satoshi Goto, “Partially-Parallel LDPC Decoder Based on High-Efficiency Messaga-Passing Algorithm,” Proceedings of the 2005 International Conference on Computer Design,2005.pp503～510

[3] Jun Heo and Keith M.Chugg. “Optimization of Scaling Soft Information in Iterative Decoding Via Density Evolution Methods”, IEEE Transactions on Communications, Vol.53, No.6, June 2005

[4] Zhixing Yang, Zhichu Lin. “Semi-parallel decoder design of Quasi-cyclic LDPC decode” [J]. Circuit and Application. 2006. No.2, Pp24 ~ 26

[5] Yun Chen, Xiaoyang Zeng. “VLSI design of irregular LDPC decoder meeting the DTMB standard” [J]. Communication Journal. 2007. Vol.28, No.8, Pp61 ~ 66

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 106-109

Exploring Architecture-Based Software Reliability Allocation Using a Dynamic Programming Algorithm

Hui Guan[1,2], Tingmei Wang[3], and Weiru Chen[1]
[1] Shenyang Institute of Chemical Technology, Shenyang, China
[2] STRL, De Montfort University, Leicester, England
Email: guanh1999@126.com
[3] Beijing Union University, Beijing, China
Email:wtm9329@hotmail.com

***Abstract*—Software reliability allocation plays an important role during software product design phase, which has close relationship with software modeling and cost evaluation. We formulated an architecture-based approach for modeling software reliability optimization problem, on this basis a dynamic programming algorithm has been illustrated in this paper which can be used to allocate the reliability to each component so as to minimize the cost of designing software while meeting the desired reliability goal. The result of our experiment show an optimal or near optimal solution to the problem of selecting the component comprising the software can be obtained with lower cost..**

***Index Terms*—Architecture, Software Reliability, Reliability Allocation, Dynamic Programming**

I. INTRODUCTION

The impact of software structure on its reliability and correctness was highlighted as early as 1975-76 [1, 2]. However, with the rapid expansion of software system size and complexity, the software structure has shifted from the earlier structure- oriented systems to object-oriented one of the present and of the future, which has also triggered a number of efforts in development of various techniques for evaluation of these systems.

The area of the optimization of reliability allocation and development cost has gained more and more attention. Various techniques had been used in the past for designing systems under dual and often conflicting constraints of maximizing reliability and minimizing cost. The idea of software reliability allocation was first put forward by M.E. HELANDER and Niclas Ohlsson [3], they described a reliability allocation model called RCCM (Reliability Constrained Cost Minimization), which is used to assign the reliability. In addition, Zahedi and Ashrafi [4] adopt AHP method for modeling the software architecture with cost as the constraints and propose a model relating to the system reliability maximization. Boehm [5] presents a method for evaluating software development cost by using COCOMO model.[6, 7] outline a method for optimization of reliability allocation and testing schedule for a software system taking into account the reliability growth of its components.

However, the models and methods presented above are mainly applied during the late phase of software development other than the early stage. It is known that the later the defects are found, the higher cost needs to be paid for them. Software architecture is just the product of early stage in software development. If the development cost evaluation can be applied in accordance with the given software architecture and reliability requirement and be implemented before the software development, development resources can be assigned properly and well organized software development can be guaranteed under such a circumstance. The aim of this paper is to propose an idea of architecture-based software reliability allocation, address an architecture-based software reliability-cost model for evaluating the architecture before developing the software.

The discussion of this paper takes the following structure. Section 2 introduces architecture-based software reliability allocation model. Section 3 depicts how to find out the optimal allocation method by using a dynamic programming algorithm. Section 4 illustrates the application of the algorithm proposed in section 3 and section 5 offers concluding remarks and directions for future research.

II. SOFTWARE RELIABILITY ALLOCATION MODEL

A. Software development cost minimization versus reliability allocation

In fact, it is impossible to improve the software reliability while lowering the software system development cost because they are two conflicting constraints.

The so-called software development cost minimization can be considered from two points of views. One is to find an optimal reliability allocation method while achieving the given reliability such that the development cost can be as low as possible; the other is how to allocate the reliability to each component on the premise of the given cost so that the system reliability can be maximized. This paper focuses on the former one.

Many systems are implemented by using a set of interconnected subsystems. Reliability allocation means adjusting the reliability among different subsystems so that the total system development cost (including human, material resources, development time and testing time etc) can be minimized. Reliability allocation can be used to

deal with such kind of problem that the goal is set prior to the solution. Usually the number of the solution is more than one, as a result reliability allocation is used to deal with the optimal problem with some constraints.

B. Software reliability and cost model

It is reasonable to assume that the cost function f_i would satisfy these three conditions [8]:

- f_i is a positive definite function
- f_i is non-decreasing
- f_i increases at a higher rate for higher values of R_i

The third condition suggests that it can be very expensive to achieve the reliability value of 1. In fact for software, it has been shown that under some assumptions, it is infeasible to achieve ultra-high reliability in software [9].

In some cases, the cost function can be derived from basic considerations and is usually stated in terms of the reliability, as we will do below for software reliability.

According to the definition of software reliability [10]:

$$r(t) = e^{-\lambda t} \qquad (1)$$

Where r(t) is continuous-time system reliability, and λ is its failure rate.

From the formula above, it can be concluded that the reliability of a software is strongly dependent on the number of faults that remain in it after testing and debugging have finished; as fault decreases the probability that the software works according to its specification will increase, that is, reliability will vary inversely with faults.

According to the experiences from practical engineering, this paper takes account of the factors by assuming that software reliability and development cost satisfies the relations as below:

1) The number of faults existed in system is inversely proportional to system development cost, and the system failure rate is proportional to the number of faults. The relationship between them can be presented as:

$$E \propto 1/C \qquad (2)$$

$$\lambda \propto E \qquad (3)$$

Where C is the development cost and E is the number of system failures. Using equation (2), reliability function (1) is therefore

$$C \propto \frac{-1}{\ln r} \qquad (4)$$

2) Software development cost is proportional to the complexity and size of the software system.

Taking assumption 2) into account, the equation (4) can be expressed as:

$$C = \frac{-\alpha}{\ln r} \qquad (5)$$

Where α represents the complexity and size for developing the software system and is called reliability-cost coefficient in this paper.

In consideration of the operation in practical engineering, where usually exists some basic cost such as personnel training, development tools preparation etc. We use character β to represent the basic development cost in this paper. With the assumption above, we describe the reliability-cost model as below:

$$C(r) = \frac{-\alpha}{\ln r} + \beta \qquad (6)$$

Equation (6) states the relationship between the system development cost and the system reliability in our model. While determining the parameters α and β are more time-consuming and challenging, however, they can be obtained based on the prior experience of practical engineering. With regard to the basic indivisible software or component and relative to a specified software development group, α and β are the fixed values.

Software reliability-cost model is shown in the Fig.1, where r and C refer to the software reliability and the development cost. According to the model, it concluded that it will cost much in order to achieve a very high level reliability. From the following figure we can find out that the software with 100% reliability is impossible.

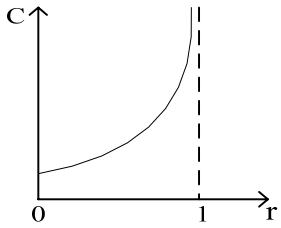


Figure 1. Reliability-cost model

Given a system with many components, the reliability of its component$_i$ can also be stated as:

$$C(r_i) = \frac{-\alpha_i}{\ln r_i} + \beta_i \qquad (7)$$

C. Software reliability allocation model

Here we assume the software system has been designed as an assembly of appropriated connected components. Let there be n components, each with reliability r_i and cost C_i, i=1…n.. Let R_{obj} be the specified target reliability and r and C represent the overall reliability and the total system cost. Let $F(r_1,r_2,\ldots r_m)$ be the function of r and r_i. The software reliability allocation model can be stated as:

Objective function

$$\min C \tag{8}$$

Subject to

$$\begin{cases} r = F(r_1, r_{2,}, ..., r_n) \geq R_{obj} \\ 0 < r_i < 1 \end{cases} \tag{9}$$

$$C = \sum C_i = \sum \beta_i - \sum \frac{\alpha_i}{\ln r_i} \tag{10}$$

The case we have discussed above is a nonlinear combination optimization problem with one object function and more than one constraint set. That is to say: try to reallocate the reliability of each subsystem or components so as to achieve the target reliability while minimizing the total development cost. Because the cost base value β of the components is nothing to do with reliability and has no impact on achieving the minimal cost, β is simplified to 0. The constraint in equation (10) can be rewriting as:

$$C = \sum C_i = -\sum \frac{\alpha_i}{\ln r_i} \tag{11}$$

As for a software system with n subsystems or components within a given architecture, the development cost of the system will increase while enhancing the software reliability target. Significantly, the greater improvement of the reliability, the more increase of the cost will be required. As shown in figure 1, the relationship between them is nonlinear.

Such a problem can be resolved with variety methods, and dynamic programming is an option.

III. USING DYNAMIC PROGRAMMING ALGORITHM TO SOLVE RELIABILITY ALLOCATION PROBLEM

Given a software system with n components and the relationship function F discussed above is known. The reliability-cost coefficient α of each component and the specified system reliability target R_{obj} is given.

The dynamic programming algorithm is as follows:

Step 1: Let S represent the reliability matrix $[r_1, r_2, ., r_n]$, T represent the cost matrix $[c_1, c_2, \ldots, c_n]$, δ be the solving step length, I_i represent the matrix with one column and n rows in which only the value of the ith element is 1 and the rest are all 0. Assume S_0=[maxr, maxr, …, maxr], maxr represents the maximized possible reliability, for example 0.9999, which means the initial reliability values of the components are all maxr.

Step 2: As for S_0, T_0 can be solved by (11), C_0 can be given by (11) and system reliability R_0 can be given by function F.

Step 3: If $R_0 < R_{obj}$ then stop and return. No solutions.

Step 4: Set Rate=0;

Step 5: for i=1 to n

i) $S' = S_0 - I_i * \delta$;

ii) With regard to S', Generate reliability R' with the function F, T' with (7), total cost C' with (11)

iii) $\Delta C = C_0 - C'$; $\Delta R = R_0 - R'$;

iv) if $R' \geq R_{obj}$ and $\Delta C/\Delta R >$Rate then Set Rate=$\Delta C/\Delta R$, R=R' , S=S' , C=C' , T=T';

Step 6: if $R_0 \neq R$ then set S_0=S; R_0=R; C_0=C; T_0=T; return to step 4

Where reliability allocation result S_0 is the reliability of each component. R_0 and C_0 are the corresponding system reliability and expected system development cost. T_0 is the expected development cost allocated to each component.

Notice from the above that prerequisite to the correctness of the algorithm is that the decrease in reliability of one component can result in that of the whole system and lower the development cost. But that can be guaranteed in our algorithm. The aim of step 5 iv) in the above algorithm is to select an optimal component whose decrease in reliability can result in the maximal cost/reliability variation, which makes the single step programming optimized so that optimal reliability allocation of the ultimate system is guaranteed.

IV. EXAMPLE AND RESULT

Here we choose a system with three independent components r_1, r_2, r_3. We assume that all the components are essential to the system and their failures are statistically independent. Therefore, the relationship between the total system reliability r and its components' reliability r_i (i=1, 2, 3) can be stated as: $r = F(r_1, r_2, r_3) = r_1 * r_2 * r_3$ Suppose that the complexities of the components are 0.30, 0.72 and 0.54 respectively. In order to minimize the system development cost and the system reliability shall be no less than 0.95, how to allocate the reliability to each component. Set the precision of computing is 0.01.

Such a problem can be rewritten as:

$R = r_1 * r_2 * r_3 \leq 0.95$

$c_1 = -0.30/\ln r_1$

$c_2 = -0.72/\ln r_2$

$c_3 = -0.54/\ln r_3$

Compute the values of parameters (r_1, r_2, r_3) with which the total cost C ($C = c_1 + c_2 + c_3$) is minimized.

With respect to each component, we compute the cost with the reliability from 0.95 to 0.99 (increment is 0.01) according to the reliability/cost function model in the data set as shown in Table 1.

TABLE I. COST AND RELIABILITY DATA SET

	r_1	c_1	r_2	c_2	r_3	c_3
1	0.95	5.85	0.95	14.04	0.95	10.53
2	0.96	7.35	0.96	17.64	0.96	13.23
3	0.97	9.85	0.97	23.64	0.97	17.73
4	0.98	14.85	0.98	35.64	0.98	26.73
5	0.99	29.85	0.99	71.64	0.99	53.73

According to the algorithm above, set initial state S_0 = [0.99, 0.99, 0.99]. Accordingly, T_0 = [29.85, 71.64,

53.73], δ=0.01, and the system cost C_0= 29.85 + 71.64 + 53.73 = 155.22, system reliability R_0= 0.99 * 0.99*0.99 =0.97.

Set i=1, 2, 3, then compute separately with different value:

1) S'= S_0 – [0.01, 0, 0] = [0.98, 0.99, 0.99], R'=0.96, T'= [14.85, 71.64, 53.73], C'=140.22

ΔC= 15, ΔR=0.01, ΔC/ΔR=1500

2) S'= S_0 – [0, 0.01, 0] = [0.99, 0.98, 0.99], R'=0.96, T'= [29.85, 35.64, 53.73], C'=119.22

ΔC= 36, ΔR=0.01, ΔC/ΔR=3600

3) S'= S_0 – [0, 0, 0.01] = [0.99, 0.99, 0.98], R'=0.96, T'= [29.85, 71.64, 26.73], C'=128.22

ΔC= 27; ΔR=0.01, ΔC/ΔR=2700

Choose the optimal result 2), set S_0= [0.99, 0.98, 0.99], continue to perform the same operation :

1) S'= S_0 – [0.01, 0, 0] = [0.98, 0.98, 0.99], R'=0.95, T'= [14.85, 35.64, 53.73], C'=104.22

ΔC= 15, ΔR=0.01, ΔC/ΔR=1500

2) S'= S_0 – [0, 0.01, 0] = [0.99, 0.97, 0.99], R'=0.95, T'= [29.85, 23.64, 53.73], C'=107.22

ΔC= 12, ΔR=0.01, ΔC/ΔR=1200

3) S'= S_0 – [0, 0, 0.01] = [0.99, 0.98, 0.98], R'=0.95, T'= [29.85, 35.64, 26.73], C'=92.22

ΔC= 27; ΔR=0.01, ΔC/ΔR=2700

Choose the optimal result 3), set S_0= [0.99, 0.98, 0.98], and then continue to perform the same operation, all of the results R' are less than the specified reliability target 0.95. Therefore, the reliability allocation in this case is as below:

1) System reliability allocation S_0= [0.99, 0.98, 0.98];

2) System reliability R_0=0.95;

3) Expected system development cost C_0= 92.22;

4) Expected development cost assigned to each components T_0 = [29.85, 35.64, 26.73].

V. CONCLUSION AND FUTURE WORK

Software reliability allocation plays an important role during software product design, which has close relationship with software modeling and cost evaluation [11]. We formulated an architecture-based approach for modeling software reliability optimization problem, on this basis a dynamic programming algorithm has been illustrated in this paper which can be used to allocate the reliability to each component so as to minimize the cost of designing software while meeting the desired reliability goal. The result of our experiment show an optimal or approximate optimal solution to the problem of selecting the component comprising a software can be obtained with lower cost (a high reliability). The reliability and cost allocation model presented in this paper can be used to solve the optimal allocation problems in simple systems, it is also applicable in complex systems.

We did not consider how to set the value of solving step length δ in this paper. However, the greater of δ will result in the imprecise solving result and the smaller of δ will slow down the speed of solving. We can adopt a technique of variable step length relating to solving precision to resolve such problem. Another problem is to do with setting the initial value for S_0. In our algorithm, we suppose a reliability value *maxr* with maximal possibility. But how much should the value be? Whether should an initial value be set in favor of solving more quickly? These problems deserves further studied.

REFERENCES

[1] D.L.Parnas. "The Influence of Software Structure on Reliability". In *Proc.1975 Int'l Conf. Reliability software*, Los Angeles, CA, April 1975. pp. 358-362.

[2] M.L.Shooman. "Structural models for software reliability prediction". In *Proc. 2nd Int'l Conf. Software Engineering,* San Fransisco, CA, October 1976, pp. 268-280.

[3] M.E. HELANDER, M. Zhao and N. Ohlsson. "Planning Models for Software Reliability and Cost". *IEEE Trans. on Software Engineering,* 1998, 24(6):420~434.

[4] F. Zahedi and N. Ashrafi, "Software Reliability Allocation Based on Structure , Utility , Price and Cost ". *IEEE Trans. on Software Engineering,* 1991, 17 (4):345 - 356.

[5] B. Boehm , R. Valerdi , J A. Lane et al, *COCOMO Suite Methodology and Evolution*, CrossTalk, 2005, pp. 20 - 25.

[6] C. Y. Huang, J. H. Lo and S Y. Kuo, "Optimal Allocation of Testing resource Considering Cost, Reliability, and Testing Effort", In *Prof. 2004 Pacific Rim Dependable Computing*, French Polynesia, 2004, pp.103 - 112.

[7] S. Y. Kuo, C. Y. Huang and M R. Lyu, "A Framework for Modeling Software Reliability , Using Various Testing Efforts and Fault Detection Rates". *IEEE Transactions on Reliability*, 2001, 50(3) :310 - 320.

[8] A. Mettas, Reliability allocation and optimization for complex systems. In *Proc. Annual Reliability and Maintainability Symposium,* Los Angeles, CA, January 2000,pp.216-221.

[9] R. W. Bulter and G.B. Finelli, "The infeasibility of quantifying the reliability of life-critical real-time software", *IEEE Trans. on Software Engineering*, 1993,19:3-12.

[10] M. R. Lyu. *Handbook of Software Reliability Engineering*. IEEE Computer Society Press, New York, l996, pp.36.

[11] M. R. Lyu. *Handbook of Software Reliability Engineering*. IEEE Computer Society Press, New York, l996, pp.3l5

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 110-113

Performance Analysis of IP over Hierarchical WDM Ring Networks

JihHsin Ho
Department of Information Management, Diwan University, Tainan 721, Taiwan, ROC
E-mail: hjsin@dwu.edu.tw

***Abstract*—Wavelength division multiplexing (WDM) appears to be the solution of choice for providing a faster networking infrastructure that can meet the explosive growth of the Internet. It has the scalability problem with increasing the access node numbers in WDM ring architecture, so we further propose the alternatives of hierarchical WDM ring architecture. In this architecture the number of routing nodes and user nodes shown as formulas and propose the routing rules, using CSMA/ID network protocol avoid packet collision. Finally present performance analysis of hierarchical WDM ring and compared with previous WDM single-ring network.**

***Index Terms*— Scalability、Hierarchical WDM Ring Network、Routing Rules、Performance Analysis**

I. INTRODUCTION

With the explosion of information traffic due to the Internet, electronic commerce, computer networks, voice, data, and video, the need for a transmission medium with the bandwidth capabilities for handling such a vast amount of information is paramount. Recently, the channel bandwidth of commercial WDM (Wavelength Division Multiplexing) communication systems has reached to OC-192 (10 Gbps)[1], and the total bandwidth of an optical fiber exceeds 1 Tbps. This indicates that WDM is the solution for bandwidth insatiability.

However, harnessing this unprecedented bandwidth in the metropolitan network environment will require a WDM transmission protocol to efficiently transport IP (variable length) packets across the data centric WDM-based MANs. Due to the rapidly increasing services and user population on the Internet, IP packet traffic dominates the utilization of data networks. However, such packets are now transferred, switched, and manipulated through complex protocol stacks, such as IP/ATM/SONET/WDM, IP/SONET/WDM, etc. Thus, the goal of merging and collapsing the middle layers of these stacks to reduce cost, complexity, and redundancy has become an important research issue. In order to minimize the layering complexity and costs of SONET and ATM, the packet-based network traffic should be accommodated directly on the WDM network, which would to be an efficient and economical way to implement the next generation of the Internet. In this way, both the equipment cost and the management complexity related to electronic multi-layer solutions are significantly reduced in all-optical IP-over-WDM networks.

To increase network bandwidth and decrease packet transmission delay are current important studies due to the internet applications are popular and the bandwidth demands are raised, which can improve effectively by add fiber channels and switch speed techniques. Moreover using the hierarchical networks decrease the diameters and average distances are worth to study topics, especially the networks exceed the thousand nodes. The hierarchical ring network architectures can decrease the network diameters [2-4].

Hierarchical Wavelength Division Multiplexing Ring Networks (HWDM-Ring) are ring network architectures connected by fibers. Figure1 shows a diagram of three-level HWDM-Ring where all nodes connected two different rings which are connected rings and local rings. Processing nodes (Access nodes) connected by local rings which are connected by connected rings, and multiple connected rings are connected by higher level connected rings that achieve the hierarchical architectures. Routing nodes (Core nodes) are communicational interface connected between different level rings. Local rings are composed some processing node and one routing node and connected rings are composed all routing nodes. For example HWDM-Ring shown in figure 1 is formed by two level connected rings and one level local ring and one ring has 13 processing nodes and whole three-level HWDM-ring has 234 processing nodes. The routing nodes or processing nodes in the same connected rings or local rings own different wavelength using the WDM access technology.

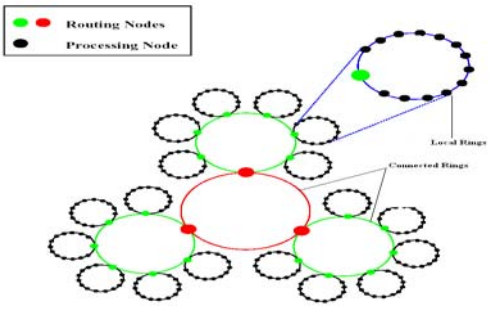


Figure 1. The Example of Hierarchical WDM Ring Networks

Although WDM provides multiple data paths, thereby reducing communication contentions, IP packet is being transmitted to a target channel while the node is detecting another IP packet arriving on the same channel at its input, an access collision will occur. A Media Access Control (MAC) protocol is necessary to prevent this phenomenon.

There are some hierarchical WDM ring networks about MAC protocol as follows [5-7]: Under TDMA (Time Division Multiple Access) each cycle is divided into time slots and each node is assigned s slot, in turn, during it has exclusive access to a channel for transmission. The length of TDMA cycle is therefore determined by length of each slot and by the number of slots needed. Under DMON each node may transmit on a data channel once it has reserved access on the control channel, but controlled by a token on a dedicated channel. The procedure a node follows to transmit a data or broadcast packet is as follows: 1) node acquires token, 2) node transmits slot reservation on the control channel, 3) node releases token, and 4) node transmits data on the reserved channel. Under THORN (Token Hierarchical Optical Ring Network) that is a variation of the DMON protocols. The procedure a node follows to transmit a data packet is as follows: 1) node acquires the necessary token on the token channel, if the token can't be acquired immediately by setting the corresponding request bit, 2) Node transmits on the channel for the token was acquired, and 3) node releases the token.

The rest of this paper is organized as follows. The hierarchical WDM ring architectures are presented in Section 2. The node architectures present in Section 3. In Section 4 introduces the CSMA/ID protocol and frame format. In Section 5, an approximate queue model for this protocol is presented to evaluate the performance. In Section 6, the numerical results are obtained from our analysis. Concluding remarks are made in Section 7.

II. HIERARCHICAL WDM RING ARCHITECTURE

A hierarchical WDM ring networks are comprised a number of levels which included many ring networks. A higher level rings are connected by routing node in the lower level rings. The processing nodes are located on lowest level and every ring has a routing node exactly connected by upper level ring besides the highest ring.

Using HWDM-Ring(B_1,B_2,…B_i, …,B_m) to present a m-level hierarchical ring included $\prod_{i-1}^{m} Bi$ processing nodes and $\sum_{i=1}^{m-1}\coprod_{j=1}^{i} B_j$ routing nodes. Take figure 1 for example, this diagram presents HWDM-Ring(3,6,13) which is 3-level hierarchical ring and included 234 processing nodes and 21 routing nodes.

Network diameter is defined as the maximum distance between the nodes arbitrarily. In the WDM single-ring network which has n nodes and the diameter is n-1. In the HWDM-Ring(B_1,B_2,…B_i, …,B_m) network, the diameter is $(B_1 - 1) + 2 \cdot \sum_{i=2}^{m} B_i$.

Giving the HWDM-Ring(B_1,B_2,…B_i,…,B_m) network which every node has a m-tuple address A(a_1,a_2, …,a_k, …a_m) means this node address located on this ring. The routing node address (a_1,a_2, …,a_k,0,0,…0) stands for the composed the source node address (a_1,a_2,…,a_k,1,0,…0), (a_1,a_2,…,a_k,2,0,…0),…(a_1,a_2, …,a_k,B_{k+1},0,…0). Giving the source node address S(s_1,s_2,…,s_m) and the destination node address T(t_1,t_2,…,t_m). The routing rule finds the path to transmit data packet from source node to the destination node. In this routing rule finds the routing node from the source node, to find the lower level routing node once more and search the highest level routing node. Inversely from the highest level routing node to find the destination node located on lowest level. For example, the 3-level hierarchical ring has source node address (0,2,6) and destination node address (2,4,9), we find the routing path using the routing rule is follow by (0,2,6), (0,2,7), (0,2,8), (0,2,9), (0,2,10), (0,2,11), (0,2,12), (0,2,13), (0,2,0), (0,3,0), (0,4,0), (0,5,0), (0,6,0), (0,0,0), (1,0,0), (2,0,0), (2,1,0), (2,2,0), (2,3,0), (2,4,0), (2,4,1), (2,4,2), (2,4,3), (2,4,4), (2,4,5), (2,4,6), (2,4,7), (2,4,8) and (2,4,9).

III. NODE ARCHITECTURE

The node included processing and routing nodes architecture of the network is shown in Figure 2. Each node has one tunable transmitter and *W* fixed receivers with one for each data channel. For the optical signal sent from upstream nodes, a splitter is used to tap off a small portion of the optical power from the ring to the receivers. Every receiver detects the optical signal carried in its corresponding wavelength within the output branch from the splitter for node address identification. If the destination address in the incoming packet header matches the node address, the packet data is sent to the host. Meanwhile, the MAC control scheme is signaled to activate the opening of the on-off switch for the corresponding data channel in order to remove the received packet carried in the major portion of the optical signal through the delay line. If the packet is not destined to this node, the detected packet is ignored and the process of scanning next the new packet is started.

IV. CSMA/ID MAC PROTOCOL AND FRAME FORMAT

To avoid packet collisions and to use bandwidth more efficiently, this paper proposes a novel Medium Access Control (MAC) protocol, named CSMA/ID, that is based on the Carrier Sense Multiple Access and Idle Detection schemes. The downstream access point recognizes the incomplete IP packet by the presence of the sub-carrier signal and pulls it off the ring. The carrier-sense can check the ACL (available channel length) to notify the Tx transmit the packet to the queue packet. Based on the protocol, each node monitors the wavelengths and detects the corresponding ACL provided that there are IP packets for transmission. Given that an IP packet is being transmitted to a target channel while the node is detecting another IP packet arriving on the same channel at its input, a dilemma of ring access (an access collision) will occur. Such collisions are due to the fact that the node cannot know if the opening is long enough to accommodate the packet. With the carrier access scheme, to guarantee the correctness of the protocol operations, the delay line inside the nodes must be used to delay the incoming packet. In addition, the delay line should be

long enough to cover the maximum IP packet length (1500bytes) so that unnecessary fragmentation can be avoiding along with packet collision and thus improve the utilization of the bandwidth. Furthermore, the fiber delay line inside the AP is responsible for processing IP packets time. The MAC protocol decides whether packet in the queue can transmit or not according to idle channel messages, transmit packet lengths and the transmission algorithms. There are three features in this protocol. First, it is a fully distributed, asynchronous protocol that does not need a centralized controller or a separate control channel to harmonize and synchronize the operations of nodes. Second, the transmitting packet will not happen collide with incoming packet on same wavelength, because the FDL length (1500bytes) is long enough. Third, it supports variable-length IP packets without complicated segmentation and reassembly, which becomes harder as the line speed of optical wavelengths increases.

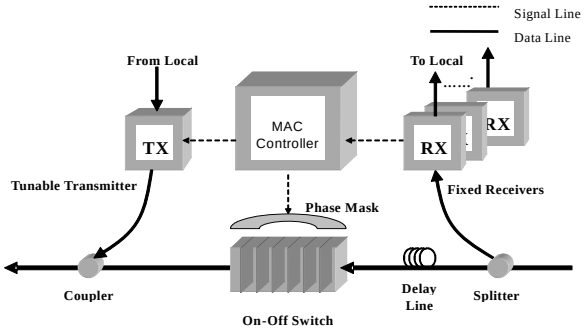


Figure 2. Node Architecture

To support the carrier access scheme, the frame format adopted is shown in Figure 3. The carrier sensing mechanism for finding transmitted packets in an optical fiber can be based on sub-carrier signaling or receiver monitoring. For sub-carrier signaling, each wavelength is associated with a sub-carrier frequency. When a node transmits a packet, it multiplexes the corresponding sub-carrier frequency. The nodes determine the occupancy of all wavelengths in parallel by monitoring the sub-carriers in the RF domain. In addition, since each receiver extract the optical signals from the corresponding data channel (or wavelength), receiver monitoring can be another approach to determine the occupancy of all wavelengths. It seems natural that the receivers are associated with the auxiliary function to monitor the status of the optical ring network. Nowadays, the cost of such receivers is still so high that they not economical to manufacture, but a cheaper process may be realized later. The start delimiter (SD) and the end delimiter (ED) mark a physical data frame conveyed in data channels for packets. The source address (SA) and the destination address (DA) serve as the address information in the network. To prevent possible transmission errors midway, the cyclic redundancy check (CRC) is employed. The flag (FG) field has the routing rule functions.

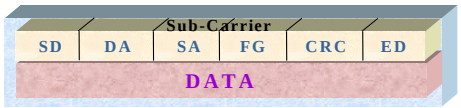


Figure 3. The Frame Format

V. NETWORK PERFORMANCE

In order to analyze the m-level hierarchical WDM ring networks, it is assumed that the bridge traffic load from the upstream source is equally distributed among W rings. To simplify the analysis, let the circulation of slots on W rings be synchronized. That is, a node can observe W MTU on different rings at the same time. Since the bridge traffic load from the upstream source is uniformly distributed among the W rings, the average bridge traffic load of each ring, ρ_B, can be expressed as:

$$\rho_B = \rho_{Bi} / W \tag{1}$$

The probability that the packet at the head of a queue cannot get an empty MTU among the currently passing W MTUs is $(\rho_B)^W$. Therefore, the probability that the packet has to wait i MTUs before it can be sent out is $(\rho_B)^{W \cdot i}(1-(\rho_B)^W)$.

Let $E[d_B]$ be the average time [8-9] required to find the arrival of an empty MTU, then we have

$$E[d_B]=\sum_{i=0}^{\infty} i\frac{L}{R}(\rho_B)^{W\cdot i}(1-(\rho_B)^W)=\frac{L\cdot(\rho_B)^W}{R\cdot(1-(\rho_B)^W)} \tag{2}$$

Using the queuing model [10] analyzes the packet average transmission time for maximum diameter. Since for each packet in the queue the arriving packet has to wait for L/R, the average queuing delay in the queue faced by an arriving packet is

$$TQ = E[\alpha] + \lambda_i TQE[X] + \lambda_i TQE[d_B] \tag{3}$$

Where TQ is average waiting time, $E[\alpha]$ is the packet residual time, λ_i is the packet arrival rate and $E[X]$ is the packet service time. Therefore, we have

$$TQ = \frac{E[\alpha]}{1-\lambda_i E[X]-\lambda_i E[d_B]} \tag{4}$$

The average transmission delay is

$$\begin{aligned} S &= E[X] + E[d_B] \\ &= E[X] + \frac{L\cdot(\rho_B)^W}{R\cdot(1-(\rho_B)^W)} \end{aligned} \tag{5}$$

Where S is packet transfer time, L is the delay line length, R is the channel speed and ρ_B is the channel traffic load. Thus, the average transfer delay for maximum diameter is given by

$$D = TQ + S + \tau/2$$

$$D = TQ + S + ((B_1 - 1) + 2\cdot\sum_{i=2}^{m} B_i)\cdot\eta, \tag{6}$$

Where η is the propagation delay between two adjacent nodes.

VI. RESULTS AND DISCUSSIONS

In this section, we will present the analytical and simulation results of packet transfer delay of the network. To evaluate and compare the performance of the hierarchical WDM ring network and WDM single-ring network, the following parameters have been listed below.

The simulation experiments are based on the codes by SIMSCRIPT II and are replicated corresponding to variance reduction technique with different sequences for

pseudo random numbers. The results are obtained with 95% confidence level.

The WDM single-Ring network:

— Number of processing nodes 20
— Number of channels 4
— Propagation delay between adjacent nodes 14μsec
— Channel speed 10 Gbps
— Size of the delay line 800 bits
— Average IP packet size 512 bytes

The hierarchical WDM ring network:

— 3-level HDWM-Ring(2,2,5) with the number of processing nodes 20 and routing nodes 6.
— Number of channels 1,2,4
— Propagation delay between adjacent nodes 14μsec
— Channel speed 10 Gbps
— Size of the delay line 800 bits
— Average IP packet size 512 bytes

Figure 4 presents the simulated and analytical results of the average packet transfer delay in this network. The curves demonstrate that a high node offered load can be achieved with low transfer delay when the number of channels is large. The agreement between the simulation results and the analytical results is excellent.

The performance metric using average transfer delay in comparison with the WDM single-ring network is shown in Figure 5. Under the steady state network condition, the average transfer delay characteristic of the network with a short diameter distance is better than that of a long diameter distance WDM ring network.

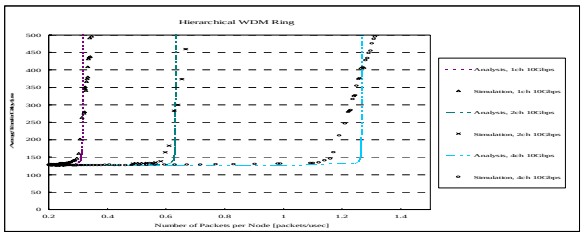


Figure 4. Average Transfer Delay for Various The Number of Channels

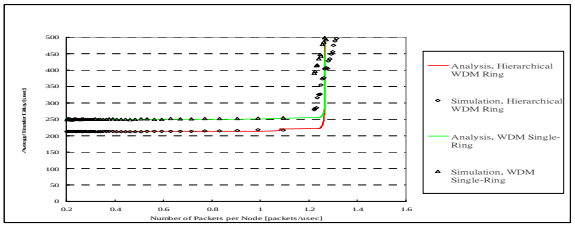


Figure5. Comparing Hierarchical WDM Ring and WDM Single-Ring

VII. CONCLUSIONS

In summary, in this paper we have investigated a hierarchical WDM ring networks. A novel MAC protocol for all optical WDM ring networks supports the transmission of IP packets directly over hierarchical WDM from LAN to MAN. This protocol can avoid packet collision, reuse wavelength and does not require a fragment packet scheme. For verification, a simulation program obtains simulated results for the network, and the simulated results closely resemble the analytical values, and this demonstrates the performance of the network. It is also observed that the transfer delay characteristic of the network is almost proportional to the number of channels and the diameter distance in the network. With regards to the utilization of bandwidth of all hierarchical networks, our protocol displays the excellent characteristics of high throughput, low delay for all optical communications.

ACKNOWLEDGMENT

The authors would like to thank the Diwan University in R.O.C. for financially supporting this research, Grant NO. DWU97C05.

REFERENCES

[1] Paul Veitch , “Resilience for IP over DWDM backbone networks”, Electronics & Communication Engineering Journal”, Feb. 2002, pp.39-48.

[2] S. Dandamudi and D. Eager. Hierarchical Interconnection Networks for Multicomputer Systems. IEEE Transactions on Computers, 39, Jun. 1990, pp. 786 – 797.

[3] P. Dowd, K. Bogineni, K. A. Aly, and J. A. Perreult. Hierarchical Scalable Photonic Architectures For High Performance Processor Interconnection. IEEE Transactions on Computers, 42, Sep. 1993, pp.1105 – 1120.

[4] A. Louri and R. Gupta. Hierarchical Optical Interconnection Network (HORN): Scalable Interconnection Network for Multiprocessors and Multicomputers. Applied Optics, 36, Jan. 1997, pp.430 – 442.

[5] K. Sivalingam and P. Dowd. A Multilevel WDM Access Protocol For An Optically Interconnected Multiprocessor System. IEEE Journal of Lightwave Technology, 13, Nov. 1995, pp.2152 – 2167.

[6] J. Spragins, J. Hammond, and D. Powlikowski. “Telecommunications: Protocols and Design.” Addison-Wesley Publishing Co., Reading, MA, 1991.

[7] Tomas S. J and Ahmed. L, “Media Access Protocols for a Scalable Optical Interconnection Network”, IEEE.

[8] J.H.Ho, W.P. Chen, W.S. Hwang and C.K. Shieh," Performance Evaluation of CSMA/ ID MAC Protocol for IP over WDM Ring Networks,"International Journal of Communication Systems, vol.21 ,no. 11, pp.1155-1170, Nov.2008.

[9] [9] W.P. Chen, J.H.Ho, W.S. Hwang and C.K. Shieh,"Novel MAC Protocol with Idle Detection for All-optical WDM Ring Networks,"Journal of Optical Networking, vol.8,no.2,pp.112-129,Feb.2009

[10] L.Kleinrock,“Priority Queueing”, Queueing System II.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 114-119

Three-Tier Security Model for E-Business: Building Trust and Security for Internet Banking Services

Yu Lasheng [1], and MUKWENDE Placide[2]
[1] Central South University/Department of Computer Science, Changsha, China
Email: ley462@163.com
[2] Central South University/Department of Computer Science, Changsha, China
Email: mukwende@gmail.com

***Abstract*—The biggest problem facing Internet banking today is the thorny issues of trust and security of online transactions. In fact, the vast majority of customers are concerned about the safety of their transaction, and they can't simply trust the web fearing that their transactions and credentials might not be safe due to the increasing number of online Internet attacks. A new model for processing Internet banking transactions is presented in this paper, it increases trust and security over the existing model, by allowing customers and banks to authenticate each other, and sign processed transactions online, It enhances security through use of a three-tier, trusted, layered, and secure channel. The model ensures that only qualified people can access Internet banking accounts, that the information viewed remains private and can't be modified by third parties, and that any transactions made are traceable and verifiable.**

***Index Terms*—Internet banking, security model, transaction signing, mutual authentication, Application Layer Security**

I. INTRODUCTION

The emergence of the Internet as the global distribution medium is motivating the banking industry to grow their computerized network through the use of Internet Banking. Doing such business via Internet introduces new challenges for security and trustworthiness. Trust and security are key enablers of the Information Society; specifically, they are the first and foremost requirements needed to be addressed by Internet banking systems. For customers to use Internet banking services comfortably, they must have confidence that their online services are trustworthy and secure. Similarly, for banks to provide Internet banking services they need confidence in the security of online transactions.

Internet security is well known and many security models and protocols have been developed for it. Secure Sockets Layer/Transport Layer Security (SSL/TLS) is recognized as the de facto Internet banking standard to offer trust and security for transactions [1]. It is claimed by Certification Authority(s) that the use SSL Certificate on company's Web server can securely collect customer's sensitive information online, win customer's trust, and increase business by giving customers confidence that their credentials and transactions are safe [2]. However, nowadays, trust and security has been diminished by the increasing number of local attacks (malicious software on client side such as, Trojan-horse), remote attacks (phishing, pharming), which are used to steal customer's credentials or SSL user session. An attacker can combine local and remote attacks; this can result in more serious damage [3].

In traditional banking, trust and security results from: firstly, customer and Bank to authenticate each other; secondly, they conducting transactions in a secure environment; and finally, signing and keeping copies of the transaction sheets by either party. This paper uses the same approach to restore trust in a digital environment by authenticating bank and customer using physical credentials to access Internet banking accounts, a three-tier security model is used to provide a secure environment. Digital signatures are also used to imitate traditional paper-based signature into the digital realm by adding a digital "fingerprint" as a signature to an electronic transaction document, and either side keeps a copy of the signed document [4].

Firstly, this paper presents approaches of existing Internet banking security models, followed by the weaknesses of using SSL/TLS protocol alone to provide trusted environment for tunneling transaction data across Internet. Next, the paper shows how to create a trusted and secure environment using layered security protocols, and then proceeds by creating a Challenge-Response authentication scheme which is used to authenticate both customer and the bank. Finally we design an overview of a transaction singing scheme which shows how either party is signing the transaction document and how each one keeps a copy of the document, effectiveness of the new model are revealed. We end-up by highlighting what is important about this paper and possible considerations for future researchers.

II. EXISTING INTERNET BANKING MODELS

Internet-based (electronic) banking schemes rely on the existence of an Internet connection over which a customer can access bank services [5]. Customers can use existing "browser" software such as Mozilla Firefox or Microsoft's Internet Explorer as the client interface to the bank system. In this model, the bank's server provides

HTML forms-based interface through which customers can make requests and conduct transactions, communication security is provided by the SSL protocol which is built into the browser, or else, Customers can download Java applets from the bank-server's web site. The downloaded applet provides the interface through which customer transactions can take place. In this case, communication security is provided by the applet in addition to the security provided by SSL [E] (see Fig. 1).

Any Internet banking system must solve the issues of authentication, confidentiality, integrity, and nonrepudiation; to ensure that only qualified people can access Internet banking accounts, that the information viewed remains private and can't be modified by third parties, and that any transactions made are traceable and verifiable [1]. For confidentiality and integrity, SSL/TLS is the de facto Internet banking standard, whereas for authentication and nonrepudiation, no single scheme has become predominant yet [1]. For that reason, the diversity of Internet banking models which exist today (using SSL as the trusted tunnel) depends on authentication methods available, and security level of a model depends on authentication mechanism used to counter attacks. Specifically, Internet banking authentication methods are classified according to their resistance to two types of common attacks: *offline credential-stealing attacks, and online channel-breaking attacks*. Figure 2 shows security level of existing models.

Offline credential stealing attacks aim to fraudulently gather a user's credentials either by invading an insufficiently protected client PC via malicious software (such as a virus or Trojan horse) or by tricking a user into voluntarily revealing his or her credentials via phishing. Online channel-breaking attacks, instead of trying to get the user's credentials, the intruder unnoticeably interrupts messages between the client PC and the banking server by masquerading as the server to the client and vice versa [1]. The level of security (see Figure 2) depends on whether crosses the offline-credential stealing attacks boundary (in horizontal direction) and/or online-channel-breaking attacks boundary (in vertical direction).

The Public Key Infrastructure (PKI) hard tokens, using challenge based authentication, has been proved to cross both attacks-boundaries [3] [1]. However, nonrepudiation

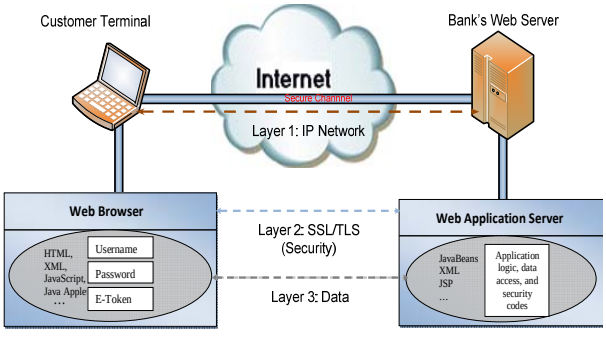


Figure1. Single-layer SSL web-based Security model for Internet banking system

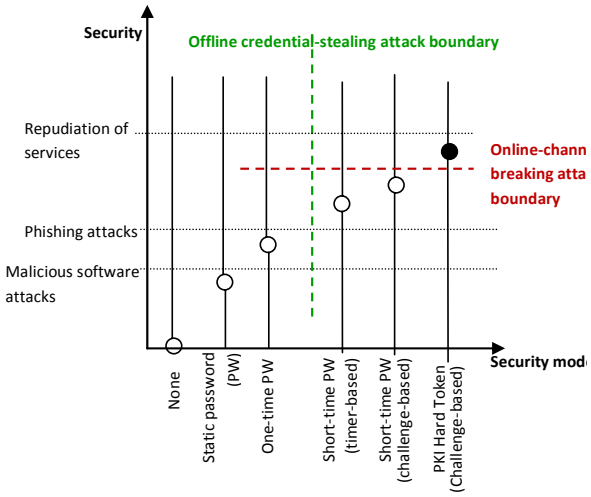


Figure 2. Comparison of existing models based on authentication mechanisms, using SSL tunneling

of transactions by any participating party (bank or customer) cannot be achieved. In addition due to limited knowledge of users to assess the difference between fake and authentic servers, secure and non-secure servers, protected and non-protected clients; reduces the level of security and trust provided by SSL tunnel, which implies the requirement to strengthen this tunnel with other tunnels.

III. Challenges of Single-Layer-Based Security Models Using SSL/TLS Protocol

SSL/TLS protocol being used as the de-facto Internet security standard; provides authentication, confidentiality, integrity and nonrepudiation of messages transmitted over Internet between the web browser and the web server only [8] [11]. However, this protocol operates below the Application layer in TCP/IP networks and doesn't provide way to ensure whether a user is, in fact, who he or she claims to be by asking for direct or indirect proof of the knowledge about some sort of secrecy or credential.

It is a common mistake for some users to believe that their online banking sessions are perfectly safe when they use an SSL connection. Security experts continually state that everything is safe if there is a yellow padlock symbol in the browser window or the URL start with *"https"* rather than *"http"* [2]. The following facts explain why SSL doesn't guarantee the safety and security of transaction over the Internet:

– SSL is designed as a secure tunnel from the end user computer's browser to the server's web server of the bank, doesn't protect the end points such as user's computer. A Trojan exploits this security hole. In addition to that, SSL is beyond providing end-user authentication services [3] [6].

– The security offered by SSL is based on the use of digital certificates of financial entities' web servers for which many internet users are not able to discern the validity of a certificate, and may not even pay attention to it [6].

– Different browsers versions will offer different levels of security as some are restricted to the use of strong cryptography. For example, some older versions of Netscape and Internet Explorer will even be restricted to offering only weak encryption, unless they are connecting to servers using Server-Gated Cryptography enabled SSL certificate. So, depending on the browser's vender and version some will only be capable of encrypting at 40 or 56-bit encryption, while more recent browser versions are capable of 128 and even 256-bit encryption key [2].
– Not all Certification Authorities may be validated by all browsers. Some are recognized by a number of browsers, and there are even increasing number of fake CAs which may be recognized by some browsers [2]. These limit the service portability as some banks are enforcing security by restricting customers to use a particular browser (for instance: Bank of China restrict customers to use Internet Explorer, which implies that customers wanting Internet banking services are requested to use only Microsoft operating systems).

The above mentioned facts prove that SSL is not enough to provide trust and security required for Internet banking. Therefore, it is mandatory to add other security protocols below and above it in order to provide a trusted environment for customers and banks to process their transactions safely.

IV. Building Trust and Security with a Three-Tier Security Model

Trust and security are key enablers of the Information Society; specifically, they are the first and foremost requirements need to be addressed by Internet banking systems. For customers to use and feel comfortable with Internet Banking services they must have confidence that their online services are trustworthy and secure. Similarly, for Banks to provide Internet banking services they need confidence in the security of online transactions.

Trust and security are very closely connected. Trust depends on the actual architecture of the security management system, but the bottom line to gain users' trust, the security management system must ensure users that the system is secured and well-protected [7]. In traditional banking trust and security are built-up by many reasons; but, the most important being: First, every bank must be authorized and certified by the controlling government to issue banking services; and then, customers are also certified by the government; banks process services which they offer in a secure environment (which is a secure office); next, customers requesting services need to authenticate themselves to the bank, similarly, banks are authenticated customers before starting their transactions; finally, each party verifies, validates and signs transaction documents, and keeps copy of the signed document.

"Authentication + Encryption + Certification Authority = Trust" [2].

Authentication, Encryption, and Certification Authority are well known security mechanisms for processing Internet-based services for a very long time, and they are currently in use by the existing Internet banking models. However, they are not able to provide the required level of trust and security (for Internet banking) depending on the way they are used. In this section we explain how to adapt the traditional banking approach to increase the level of trust and security (over the existing one-tier SSL-based security model) using three-tier model for Internet banking.

A. Building a Secure Environment

In computing industry services reliability is achieved through duplication of all services involved over a number of different service providers at different levels. Thus, reliable security can be achieved by duplicating it over different levels. The existing environment (SSL trusted tunnel) has proved to have some weaknesses, and the level of security depends on the authentication mechanism used (back to section 2, Figure 2). Therefore, we need, first, to examine why high security provided by SSL is being weakened, and then try to build a secure environment by taking into account those weaknesses. SSL is designed as a secure tunnel from end-user's web-browser (on client host) to the bank's web-server (on server host). A Trojan exploits this fact. For instance: a Trojan which drops a DLL and registers its CLSID as a browser helper object in the registry, is able to intercept any information that is entered into a web page before it is encrypted by SSL and sent out. (This is an example of a credential stealing attack). Another example: a Trojan running on an infected computer can alter the local host's file to redirect any request to an IP address controlled by the attacker. The Trojan can also install a self-signed root certificate on the infected computer (using free tools like OpenSSL to create these certificates), this enables attacker to generate official-looking SSL connections from the infected computer to the malicious web server hosting the spoofed Internet banking application. Once the user has been trapped on such a spoofed (fake) Internet banking application, the attacker can act as man-in-the-middle and relay any challenge-response protocol that might be implemented by the original Internet banking application system. This in an example of a channel-breaking attack using a malicious software (a Trojan). Thus, SSL by itself is neither able to pass the offline-credential stealing attacks boundary nor able to cross the channel-breaking boundary. The only guaranteed way to counter against any form of channel-breaking attack: is by carefully checking the IP addresses

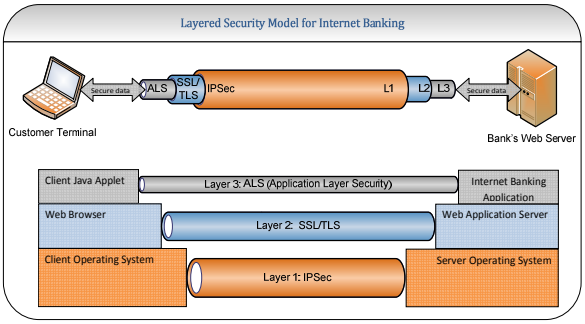


Figure 3. Three-tier Security Model for Internet banking

involved in the session and their owners. To counter against all forms of credential-stealing attacks, is by: extending the number of passwords of password to infinity; encrypt the password before it is entered; and finally, change the method for entering authentication data, such that password can never be intercepted.

Therefore, our model (see Figure 3) provides trustworthy security for Internet banking, by reinforcing SSL tunnel with two tunnels given one at the IP layer, and another at the Application layer, of TCP/IP networks.

1. Security Layer 1: Internet Protocol Security (IPSec Tunnel)

Internet Protocol security (IPSec) is a protocol, not a service, that provides confidentiality, integrity, and authentication services for IP-based network traffic. [8] Because IPSec provides host-to-host protection, can be used to counter network threats, including eavesdropping, tampering, man-in-the-middle attacks, IP spoofing, and other password-based attacks [11]. Taking advantages of IPv6 which has built-in support services for confidentiality and integrity of messages between hosts' operating systems, with its huge address space, Internet Service providers will have sufficient IP addresses to allocate enough addresses to every customer so that every IP device has a truly unique address—whether it is behind a firewall or not. In that situation where a customer is having a computer with a fixed IP address from which he always performs Internet banking operations; it is desirable to configure the bank's web server and customer's client to always establish a Virtual Private Network before customer's access to his Internet banking account. The recorded IP will always be used to authenticate the client host and its owner, and this will thwart all forms of channel-breaking attacks.

2. Security layer 3: Application Layer Security (ALS)

Application layer security refers to methods of protecting web applications at the application layer from malicious attacks that may expose private information—counter against all form of credential-stealing attacks [9] [10]. We achieve these by using tamper resisting offline smartcards that are based on Public Key Infrastructure (PKI): credentials are encrypted using public-private keys encryption before they are entered on the client computer (which may be infected). The number of password is increased from one to infinity using Challenge-based Mutual Authentication mechanism—user can choose any random challenge number at any time (see Figure 4). To counter against any cross-site scripting attacks, symmetric encryption is done with the use of secret key (K_S) between the client Java Applet (*A*) and bank's server (*B*). It is necessary to note that the secret key needs not to be encrypted because of the security provided by the lower level layers. Using ALS above SSL, increases further the required level of confidentiality and integrity required for Internet banking. However, readers need to know that ALS is not a standard protocol as there is no single scheme that has become predominant yet for it [10].

B. Challenge-Based Mutual Authentication

In the age of faceless Internet banking, authentication provides crucial online identity; customers and banks need to get to know one another before conducting business. The use of SSL to authenticate the bank's server to customer is weakened by the use of digital certificate which the customer is not able to discern, particularly, SSL doesn't provide user authentication which leads to remote attacks such as phishing, pharming and password-guessing. Use of simple passwords, even one-time passwords as well as token based authentication is vulnerable to local attacks such as Trojan-horse. A better way to provide authentication in order to improve security is through use of tamper resisting *PKI Smartcards to identify customers*, and *Challenge-Response protocol to authenticate the bank's server*. Figure 4 addresses how a customer and the bank authenticate each other. The bank's server is authenticated with the use of an offline smartcard, which has a built-in public key of the *B*ank (PU_B), a public and private key pair of the *C*ustomer (PU_C, PR_C) and a public encryption algorithm such as RSA. The Server is maintaining all the public keys of customers with corresponding smartcards secrete identification (ID_C) numbers.

For authentication the following steps (matched by number to figure 4) occur:

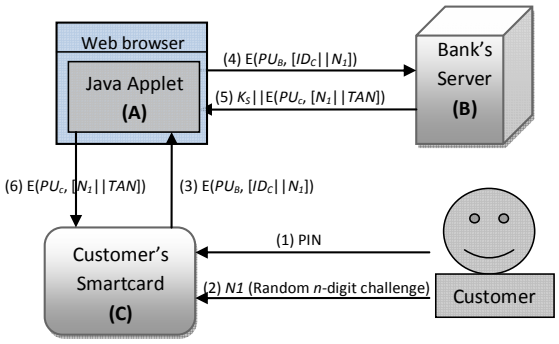


Figure 4. PKI Challenge-Response Authentication Model

1. Customer is authenticated to the smartcard by entering the *PIN* number.

2. Customer enters an *n*-digit random challenge (N_1) into the smartcard.

3. The smartcard which is used in an offline card reader uses server's public key to encrypt customer's identity (ID_C) and a random challenge (N_1). The encrypted message is sent manually to the Java *A*pplet client (*A*).

4. *A* forward the encrypted message to the bank's server (*B*). *B* decrypts the message with his private key and verifies the ID_C and recovers N_1.

5. *B* sends a message to *A* containing session key (K_S) and a message encrypted with customer's public key

(PU_C) and containing customer's challenge number (N_1) as well as a transaction number (*TAN*).

6. The message received by *A* is entered manually into the smartcard (*C*), *N1* is recovered and *TAN* is stored for future use during this session.

The ID_C must be a secret code that uniquely identifies the customer and his credit card. This code is sent to the server to authenticate the customer. For example; ID_C may be a function of anything that can be used to identify customer (such as Social Security Number or username) and password (*PIN*), $ID_C = f_{k1}(SSN, PIN)$, and this value must be kept in banks database as $f_{k2}(SSN, PIN)$ where *k1* and *k2* two different secret keys for a hash function *f*. This authentication scheme leaves an insider with little information which may be used to access customers banking accounts. The scheme verifies also the client and bank's certificates, which enhances the degree of trust between customer and bank.

C. Transaction Signing

In traditional banking, trust on the performed transaction comes from involving the two parties in approving and signing the transaction's agreement document and put a stamp on it; each party keeping a copy of the signed and stamped document after the transaction is over. Clearly, no one can refute the transaction as his signature is unique, and no one else knows how to sign the document except him. The same procedure can be used in e-business where a customer uses his PKI Smartcard's private key to sign the document while the bank uses its server's private key. A secret *Transaction Number* (*TAN*) generated by the bank's server acts as a stamp of the transaction. *TAN* act as the agreement number between the user and the server to perform all the transaction in the current session. TAN may be computed as a fixed length hash value of a function which takes the Customer identity (ID_C) and Bank identity (ID_B) and the current date and time; $TAN = f(ID_C, ID_B, Date \& Time)$. *TAN* acts as a time stamp during the signing of transaction. Because, ID_C, and ID_B are kept private, it leaves the attacker with zero knowledge on how to generate this stamp.

The following steps (matched by number to figure 5) occur during transaction verification and approval:

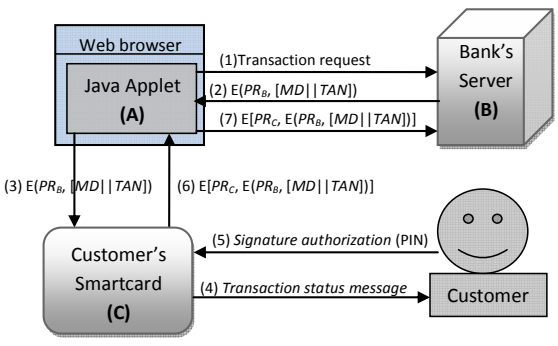


Figure 5. Transactions signing Model

1. *A* (customer's through the applet) request *B* to process the transaction.

2. *B* verifies the transaction, generates a message digest (*MD*) of the transaction, and sends to *A* a message containing *MD* and *TAN* encrypted with PR_B.

3. *A* verifies the integrity of the transaction by computing *MD* of the transaction and compare it with that one gathered from *B*. The singed message from *B* is sent to *C* for approval.

4. *C* verifies *TAN* and generates a transaction status message and requests customer to approve the transaction by entering his *PIN*.

5. Customer approves the transaction by entering the smartcard *PIN*. However, a different *PIN* may be used toward better security.

6. *C* signs the message signed by *B*, and forwards it to *A*, and stores a copy for future reference.

7. *A* forwards the message to *B*. *B* verifies the customer's signature and authorize the processing of the transaction.

V. Effectiveness of the New Model

Any Internet banking system must solve the issues of authentication, confidentiality, integrity, and nonrepudiation [1]. Confidentiality and integrity have been built-up by constructing a secure environment, (using IPSec bellow SSL, and ALS above SSL), the environment counters all forms of credential-stealing attacks as well as channel-breaking attacks: *information viewed remains private and can't be modified by third parties.* For authentication, the model extends the one-factor PKI hard tokens (challenge-based authentication) to two-factor Client-Host-IP authentication and PKI-Mutual challenge-based authentication: *only qualified people can access Internet banking accounts.* Nonrepudiation which is not provided by any of the existing models has been provided by the new model using transaction signing scheme and keeping copies of processes transaction: *any transactions made are traceable and verifiable.*

Trust is increased and ensured based on the following two hypotheses (which have been proved to be true Ref. [7]) *"Users tendency to trust is positively associated with the perceived level of security"; "Banks assurances are positively associated with the level of trust in adopting internet banking."* [7]. Figure 6, shows how the level of security has been increased with the new model. Thus, the level of trust has been increased with the new model.

The security provided by the new model which crosses the two attacks boundaries, even providing nonrepudiation service, guarantees safety which can make customers trust Internet banking. With this model, as from now, banks can guarantee customers to be confident enough that Internet banking accounts can be

accessed from customers' computer only, and that transactions made are traceable and verifiable (using even their smartcards), and if hackers attempt to do so with the use of another computer the banks will bear the liability and costs in cases.

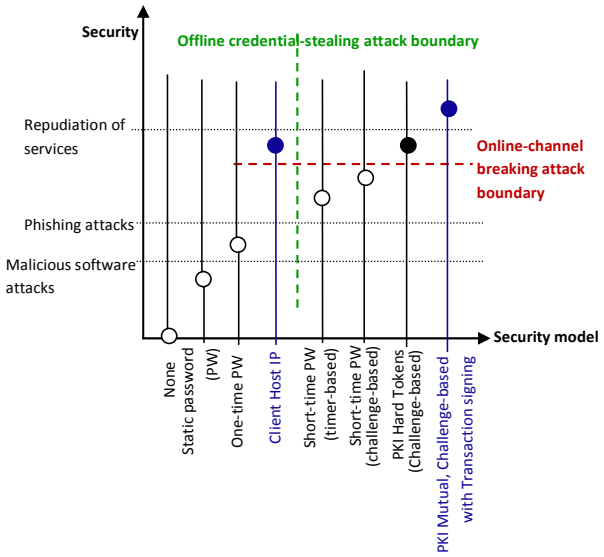


Figure 1. Effectiveness of the new model compared with the security of existing models.

VI. Conclusion

Implementing the three-tier security model for Internet Banking will offer safe Internet banking transactions that protect both the customers and banks. Customers will gain confidence that they are sending their personal information to legitimate banks' servers and not impostors, and that the privacy of their transaction and credentials is ensured during the transmission over the unsafe network. In turn, the banks and customers will receive signed transactions that either party cannot later refute as each party will have copies of signed and stamped transactions. The use of IPSec increase trust as only the customers' PCs can be used online to access customers' related accounts. The use of PKI offline smartcard readers to provide mutual authentication, ensures that only qualified people can access banking accounts. Signing transactions ensures that the transactions are traceable and verifiable. Hence, trust level is increased proportionally to the increased level of security. However, there is one overhead of manual data exchange between the offline smartcard reader and the client applet, which requires automation in future researches.

References

[1] ALAIN Hiltgen, Zurich Thorsten Kramp, and Thomas Weigold, "Secure Internet Banking Authentication" , IEEE 2006

[2] Thawte, "The value of Authentication", http://www.thawte.com, 18 July 2009.

[3] Candid Wueest: "Threats to Online Baning," Symantec Security Response, Dublin, 2006

[4] Osama Danhash, Phu Dung Le and Bala Srinivasan, "Security Analysis for Internet Banking Models", (IEEE 2007)

[5] http://www.arx.com/documents/Bank-of-Israel-Case-Study.pdf. 13 August 2009

[6] Antonio San Martino, Xavier Perramon, "Defending E-Banking Services; an Antiphishing Approach", IEEE 2008.

[7] Prof. Ali Sanayei, Ali Noroozi, "Security of Internet Banking Services and its linkage with Users' Trust", IEEE 2009

[8] Willian Stallings, "Cryptography and Network Security Principles and Practices," Fourth Edition (2006), pp. 483-562

[9] http://www.f5.com/pdf/white-papers/intelligent-layer7-protection-wp.pdf , 22 August 2009

[10] www.javvin.com/networksecurity/CommunicationSecurity.html, 18, September 2009

[11] Ross J. Anderson, "Security Engineering: *A Guide to Building Dependable Distributed Systems*," 2001, pp. 185-206

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 120-123

Entry Optimization Computation Using Simplex Algorithm Reference Trajectory Programming

Zongzhun Zheng, Yongji Wang, Fuqiang Xie , and Chuanfeng Li
Department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan, China
Email: zzzsiemen@163.com

***Abstract*—Sequential Quadratic Programming (SQP) for trajectory optimization of entry vehicle was presented. Firstly, equations of motion were normalized and an independent variable is introduced to reduce the difficulty of iterative computation. And then, optimal control problem was transformed into a nonlinear programming problem using direct collocation method. Finally, sequential quadratic programming was presented for solving the nonlinear programming problem. According to sensitivity of the initial value and long computation time spent on iteration in trajectory optimization with multi-constraints, the simplex algorithm was provided for generating a reference trajectory rapidly to satisfy all constraints absolutely. Simulation results demonstrate that the reliable algorithm can consistently achieve the desired target conditions and satisfy all constraints effectively.**

***Index Terms*—hypersonic vehicle, trajectory optimization, sequential quadratic programming, reference trajectory, simplex algorithm**

I. INTRODUCTION

Entry trajectory optimization can be described as a nonlinear optimal control problem with constraints. It is difficult to find accurate analytic solutions of optimal control variables because of complicated nonlinear model characteristic. Two classes of traditional solution methods are acknowledged for solving the optimal control problem: indirect methods and direct methods.

Indirect methods proceed by formulating the optimality conditions according to the Pontryagin's Maximum Principle and then numerically solving a two-point boundary value problem (TPBVP). The TPBVP is extremely bad conditioned due to the high sensitivity to the initial guess of the co-states and the possibility of discontinuities. On the other hand, direct methods that transform optimal control problems into parameter optimization problems and solve the problems by nonlinear programming algorithm have been intensively studied recently. Lu [1] used a FORTRAN Feasible Sequential Quadratic Programming algorithm to numerically solve the optimal control problems. Herman & Conway [2] used Legendre-Gauss-Lobatto quadrature rules for a particular choice of internal collocation points. More recently, pseudo-spectral methods [3] have been used for direct trajectory optimization.

Either direct methods or indirect methods spend long computation time on iterations to obtain optimal control solutions. Moreover, different initial reference trajectory will affect the convergence property of iterative computation. In order to search an optimum solution rapidly, initial reference trajectory should be chosen close to the best solution exceedingly. The nonlinear simplex algorithm has been implemented successfully to a large variety of problems [4]. It builds a representation of the objective function to be minimized, only comparisons of which are needed to determine the worst vertex and thus to predict the search direction. In this way, calculation costs may be significantly decreased. The implementation of the optimization procedure is less complicated than the gradient-based method, and yields a more flexible tool for complex optimization problems.

In this paper, an entry trajectory optimization strategy was proposed. Equations of motion are normalized and an independent variable is introduced for optimization to reduce the difficulty of iterative computation. Optimal control problem is transformed into a nonlinear programming problem using direct collocation method. Sequential quadratic programming is used for solving the nonlinear programming problem. According to sensitivity of the initial value and long computation time spent on iteration, Simplex algorithm is provided for searching design parameters to achieve the desired target conditions within allowable tolerances and satisfy all constraints.

II. PROBLEM DESCRIPTION

A. Entry Dynamics

Considering the rotating earth, the normalized entry equations are given in Ref. [5]. To simplify the equations and the optimization problem, the negative specific energy e is used instead of time in the motion equations.

$$e = 1/R - V^2/2 \tag{1}$$

The 3DOF point-mass dynamics are described by following dimensionless equations of motion.

$$\begin{aligned}
&\frac{\mathrm{d}R}{\mathrm{d}e} = V\sin\gamma(VD - V\varphi_{V3})^{-1}\\
&\frac{\mathrm{d}\theta}{\mathrm{d}e} = \frac{V\cos\gamma\sin\psi}{R\cos\phi}(VD - V\varphi_{V3})^{-1}\\
&\frac{\mathrm{d}\phi}{\mathrm{d}e} = \frac{V\cos\gamma\cos\psi}{R}(VD - V\varphi_{V3})^{-1}\\
&\frac{\mathrm{d}V}{\mathrm{d}e} = (-D - \frac{\sin\gamma}{R^2} + \varphi_{V3})(VD - V\varphi_{V3})^{-1}\\
&\frac{\mathrm{d}\gamma}{\mathrm{d}e} = \frac{1}{V}[L\cos\sigma + (V^2 - \frac{1}{R})\frac{\cos\gamma}{R} + \varphi_{\gamma 3} + \varphi_{\gamma 4}](VD - V\varphi_{V3})^{-1}\\
&\frac{\mathrm{d}\psi}{\mathrm{d}e} = \frac{1}{V}[\frac{L\sin\sigma}{\cos\gamma} + \frac{V^2\cos\gamma\sin\psi\tan\phi}{R} - \varphi_{\psi 3} + \varphi_{\psi 4}](VD - V\varphi_{V3})^{-1}
\end{aligned} \tag{2}$$

where R is the normalized radial distance from the center of the earth to the vehicle, the longitude and latitude are θ and ϕ, respectively, V is the normalized velocity, γ is the flight path angle, ψ is the velocity azimuth angle, the control variables are the angle of attack α and the bank angle σ. L and D represent the lift and drag accelerations.

This is a more appropriate formulation for the entry optimization problem since the initial and terminal conditions are given at determinate energy values, whereas time plays no role. When one of terminal altitude and velocity is satisfied in Eq.(2), the other one will be met spontaneously. So convergence speed is improved and difficulty of iterative computation is reduced.

B. Performance Function

This paper minimized the heat load during entry as performance function.

$$J = \int_{e_0}^{e_f} \dot{Q}(\frac{de}{d\tau})^{-1}\, de \tag{3}$$

where the state $\boldsymbol{x} = [R\ \theta\ \phi\ V\ \gamma\ \psi]^T$ and the trajectory control $\boldsymbol{u} = [\alpha\ \sigma]^T$. Usually, a nominal α -versus-*Ma* profile is available. So trajectory programming problem comes down to give the control variable σ.

C. Constraints

The process constraints of vehicle on the maximum dynamic pressure, aerodynamic acceleration, heating rate, and quasi-equilibrium glide condition are given by

$$\dot{Q} \le \dot{Q}_{\max} \tag{4}$$

$$\left| L\cos\alpha + D\sin\alpha \right| \le N_{\max} \tag{5}$$

$$q \le q_{\max} \tag{6}$$

$$L\cos\sigma + (V^2 / R - 1/R^2) \le 0 \tag{7}$$

Vehicle owns powerful maneuverability and requires terminal constraints ($R_f, \theta_f, \phi_f, V_f, \gamma_f, \psi_f$) strictly.

III. NUMERICAL SOLUTIONS

A. Direct Collocation

By the direct collocation technique, the continuous variables are represented by discrete variables, so that the optimal control problem is converted into one of constrained nonlinear parameter programming one. Simpson's rule, used in this paper, is the third-degree Gauss-Lobatto integration rule.

The negative energy is discretized into N subintervals. Within a given subinterval [e_i, e_{i+1}], the states and controls at the endpoints are optimal variables, represented by the following vector

$$\boldsymbol{Z} = \left[\boldsymbol{x}_1^T, \boldsymbol{u}_1^T, \boldsymbol{x}_2^T, \boldsymbol{u}_2^T, \cdots, \boldsymbol{x}_N^T, \boldsymbol{u}_N^T \right]$$

Simpson's system equality constraint is formulated. The following constraint has been termed the Hermite-Simpson system constraint.

$$\Delta_i = \boldsymbol{x}_{i+1} - \boldsymbol{x}_i - \frac{h_i}{6}\left(f_i + 4f_{ci} + f_{i+1} \right) \tag{8}$$

B. Sequential Quadratic Programming

The inequality constraints and equality constraints of the reentry trajectory are given in Eqs. (4) - (8). The performance function J is given in Eq. (3). All of these constitute a nonlinear programming problem which can be solved by Sequential Quadratic Programming (SQP).

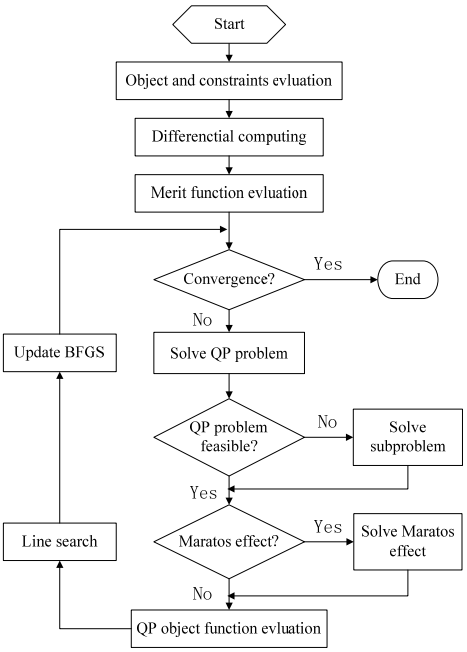


Figure 1. Flowchart of SQP method.

SQP method can be interpreted as modified Newton method for the system Kuhn-Tucker conditions (KTC) combined with some globalization scheme. SQP methods have shown enormous success for medium-scale general nonlinear programming problems. The flowchart of SQP method is show by Fig. 1.

IV. REFERENCE TRAJECTORY PROGRAMMING

A. Simplex Algorithm

The original algorithm, developed by Spendley et al [6], has been improved by Nelder & Mead [7] to enable the modification of the simplex shape in 1965. However, the original methods have been repeatedly reassessed and modified for increased robustness and the convergence rate of the algorithm. A comprehensive survey on papers that propose modified simplex algorithm with improved convergence properties is given by Kolda et al [8].

The simplex method consists in moving a regular simplex of n+1 vertices in $\mathbf{R^n}$ (a triangle in $\mathbf{R^2}$, a tetrahedron in $\mathbf{R^3}$, etc), for a problem of n parameters, each vertex representing a distinct shape. The initial simplex is built around the initial value of design variables. Then, displacements are performed in order to reduce the cost function evaluated at the worst vertex at each move. The vertex corresponding to the worst value is projected through the centroid of the remaining vertices, expecting a new shape corresponding to a better value.

Simplex algorithm is able to adapt itself to the topology while modifying its size and its shape. However,

a high stretching of the simplex may be a drawback for the convergence, if the topology changes suddenly. Therefore, an update with a regular simplex is recommended when the stretching becomes too high. As the Nelder-Mead simplex method cannot take the constraints into account, the bounds of the variation domain are implemented using barrier functions, and the eventual physical constraints may be included in the optimization through penalty functions.

B. Programming Strategy

A reference trajectory programming method is found to be more efficient in reducing search dimensions and guaranteeing fast convergence. Since the sign of σ does not impact longitudinal profile, the programming problem is decomposed into longitudinal and lateral profiles and simplified as a 3-parameter problem. Simplex algorithm is used to find appropriate design parameters.

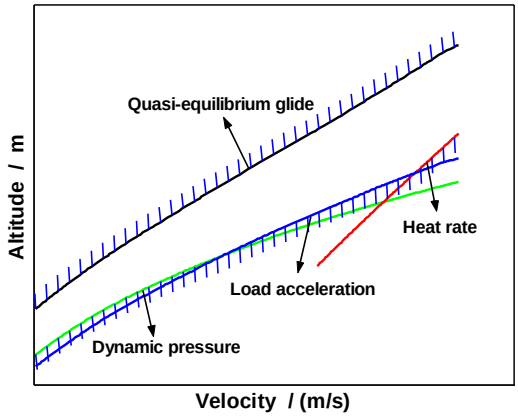


Figure 2. Altitude-velocity flight corridor.

In longitudinal profile, we design an altitude-velocity entry flight corridor firstly. Altitude-velocity corridor is described by Fig.2. Lower boundary of entry corridor makes up of heating rate curve, load acceleration curve and dynamic pressure curve. Entry trajectory in the corridor can satisfy all the mentioned process constraints consequently. Define the upper boundary $R_U(V)$ and the lower boundary $R_L(V)$. $R_L(V)$ is the envelope of the boundaries of constraints (4)-(6) after smooth processing. Choose the command altitude

$$R_C(V) = \lambda_H \cdot R_U(V) + (1-\lambda_H) \cdot R_L(V) \qquad (9)$$

with λ_H a design variable between $[0,1]$. Approaching the target, a transition from the command altitude to (V_f, R_f) is made.

The command altitude tracking can be implemented by calculating the needed lift.

$$\begin{cases} L_0 = (1/R^2 - V^2/R)\cos\gamma \\ L_C = L_0 + K_1 R_C - K_2 R - K_3 V \sin\gamma \end{cases} \qquad (10)$$

where L_0 is the needed lift by QEG, K_1, K_2 and K_3 are coefficients gained by Linear Quadratic Regulator (LQR). Then, the value of bank angle σ is given as

$$\sigma = \arccos(L_C / L) \qquad (11)$$

Conventional algorithms use heading error as a key role in bank-reversal criterion 0. The sign of the bank angle is set to the opposite of the sign of the current heading error when $|\Delta\psi| \geq \Delta\psi_{threshold}$.

The approach taken in this paper is to adopt bank-reversal twice to ensure terminal constraints, especially crossrange and heading error constraints which conventional bank-reversal criterion can not satisfy simultaneously. Two bank-reversal points e_1 and e_2, as same as λ_H in Eq.(9), are design variables of simplex algorithm. For the problem of three design parameters λ_H, e_1 and e_2, simplex is built in $\mathbf{R}^3$. Integrating equation (2) between $[e_0, e_f]$, the performance index function to be minimized is finally

$$F(x) = \sum_{i=1}^{6} K_i (y_i - y_f)^2 + \sum_{j=1}^{m} P_j [\max(g_i(x), 0)]^2 \qquad (12)$$

where $(y_i - y_f)$ is accuracy error of terminal position away demanded terminal states, K_i is weight coefficient, and P_j is penalty gene.

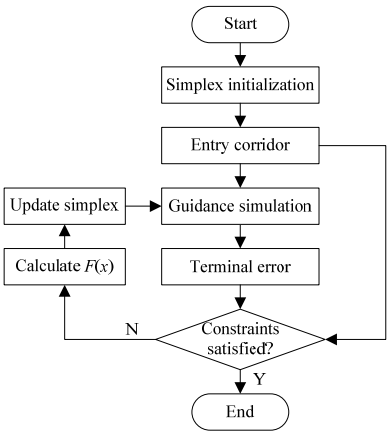


Figure 3. Flowchart of trajectory programming strategy

Simplex algorithm will be executed until all the constraints are satisfied. The flowchart of reference trajectory programming strategy is shown by Fig.3.

Ⅴ. Simulation Results

The algorithm presented in this paper is implemented in generating entry trajectory with X-33 vehicle data. The entry trajectory is constrained by $Q_{s\,max}$ = 851.169 kW/m^2, N_{max} = 2.5 g, q_{max} = 14364 Pa. The entry and terminal interface conditions are shown in TABLE I. where $H = (R-1)r_0$ is the altitude.

TABLE I. Entry and Terminal Interface Conditions

	H(km)	θ(deg)	ϕ(deg)	V(m/s)	γ(deg)	ψ(deg)
Entry	121.5	242.99	-18.26	7622	-1.4	38.239
Terminal	30.427	279.50	28.61	908.15	-7.5	41.418

First of all, reference trajectory was calculated by the simplex algorithm, based on the entry and terminal

interface conditions. And then, reference trajectory was discretized into 50 subintervals. Finally, trajectory optimization problem was solved by SQP method.

The altitude vs. velocity entry flight corridor and the reference trajectory generated by the simplex algorithm reference trajectory programming are presented in Fig.4. It is clear to see that the trajectory observes the entry corridor boundaries very well.

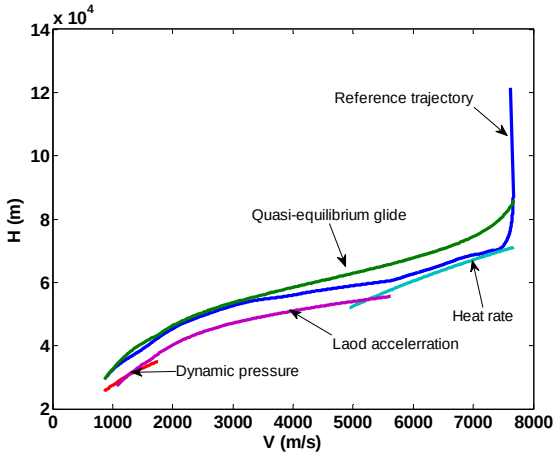


Figure 4. Entry reference trajectory

Fig.5 to Fig.6 show the optimization results by SQP. Fig.5 exhibits the bank angle vs. normalized energy and bank reversal logic. The crossrange curve is plotted in Fig. 6. It demonstrates good performance in the lateral plane toward the end.

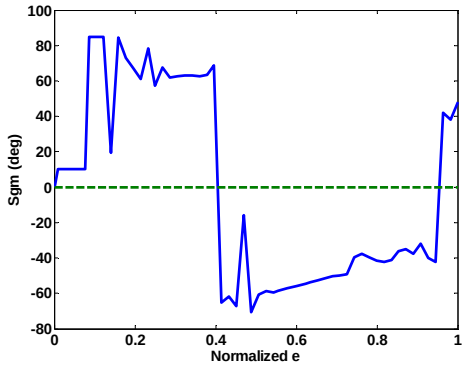


Figure 5. Bank angle command

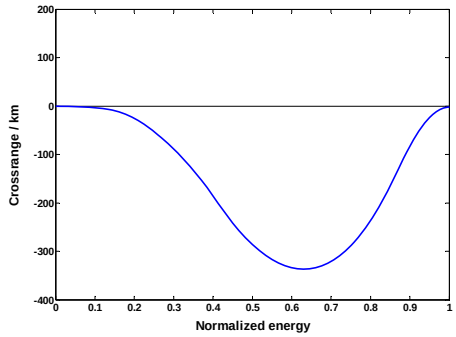


Figure 6. Crossrange curve

The terminal altitude, for which the simulation was performed, is about 30.545 km at a velocity 906.88 m/s. The longitude and latitude are 279.52 deg and 28.63 deg. The final flight path angle is -7.267 deg, with velocity azimuth angle 42.33 deg. The actual peak heat rate along the optimal trajectory was 800.12 kw/m^2, compared to the imposed limit of 851.169 kW/m^2. Normal acceleration and dynamic press were 1.61 g and 7805 Pa.

Ⅳ. Conclusions

A trajectory optimization strategy has been proposed in this paper. Equations of motion have been normalized and an independent variable has been introduced for optimization to reduce the difficulty of iterative computation. By establishing altitude-velocity flight corridor and assuming two bank-reversal points, the simplex algorithm is provided to achieve the reference trajectory and satisfy all constraints strictly. SQP was presented for solving the nonlinear programming problem. The simulation results clearly bring out the fact that all the terminal constraints are met with high accuracy and all the process constraints are satisfied absolutely. It has been shown that the desired entry performances are achieved excellently and the method is reliable.

Acknowledgment

We are grateful for the support of the National Nature Science Foundation of China, No. 60674105, the Scientific Research Cultivation Project of Ministry of Education, No. 20081383 and the 2008 Spaceflight Support Foundation.

References

[1] P. Lu, “Entry guidance for the X-33 vehicle,” *Journal of Spacecraft and Rockets*, 1998, 35(3), pp.342-349.

[2] A. L. Herman, and B. A. Conway, “Direct optimization using collocation based on high-order Gauss-Lobatto quadrature rules,” *Journal of Guidance, Control and Dynamics*, 1996, 19(3), pp.592-599.

[3] A. Benson, T. Thorvaldsen, and V. Rao, “Direct trajectory optimization and co-state estimation via an orthogonal collocation method,” *Journal of Guidance, Control and Dynamics*, 2006, 29(6), pp.1435-1440.

[4] C. Clemen., “New method for on-orbit-determination of parameters for guidance, navigation and control,” *Acta Astronautica*, 2002, 51(1-9), pp.457-465.

[5] Z. Shen, and P. Lu, “On-Board generation of three-dimensional constrained entry trajectories,” *Journal of Guidance, Control, and Dynamics*, 2003, 26(1), pp.111-121.

[6] W. Spendley, G. R. Hext, and F. R. Himsworth, “Sequential application of simplex designs in optimization and evolutionary operation,” *Technometrics*, 1962, 4, pp.441-461.

[7] J. A. Nelder, and R. Mead, “A simplex method for function minimization,” *Computer Journal*, 1965, 7(4), pp.308-313.

[8] T. G. Kolda, R. M. Lewis, and V. Torczon, “Optimization by direct search: new perspectives on some classical and modern methods”, *SIAM Reveiw*, 2003, 45(3), pp.385-482.

P. Lu, “Regulation about time-varying trajectories: precision entry guidance illustrated”, *Journal of Guidance, Control, and Dynamics*, 1999, 22(6), pp.784-790.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 124-127

Gait Recognition Based on PCA and LDA

Qiong Cheng [1], Bo Fu [2], and Hui Chen [2]
[1] School of Electrical & Electronic Engineering, Hubei University of Technology, Wuhan 430068, China
Email: qiongcheng@sina.com
[2] School of Electrical & Electronic Engineering, Hubei University of Technology, Wuhan 430068, China
Email: fubofanxx@yahoo.com.cn only4ray2002@hotmail.com

***Abstract*—This paper proposes a new gait recognition method using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). PCA is first applied to 1D time-varying distance signals derived from a sequence of silhouette images to reduce it's dimensionality. Then, LDA is performed to optimize the pattern class ificovtion..And,Spatiotemporal Correlation (STC) and Normalized Euclidean Distance (NED) are respectively used to measur the two different sequences and K nearest neighbor classification (KNN) are finally performed for recognition. The experimental results show the PCA and LDA based gait recognition algorithm is better than that based on PCA.**

***Index Terms*—gait recognition; PCA; LDA; k-Nearest Neighbor method**

I. INTRODUCTION

Gait recognition has drew a lot of attention at the leading edge of information technology researching region, since it adopts computer vision, pattern recognition, image sequence processing and other new techniques. In psychology researching [1], Johansson has indicated that MLD (Moving Light Display) could detect out pedestrian rapidly. Automatic gait recognition method, which is originally proposed by Niyogi et al., could be used to acquire gait feature from pedestrian spatiotemporal mode. Kale et al. [3] have provided a recognition method based on HMM. BenAbdelkader and some [4] make use of feature of pedestrian image autocorrelation in gait recognition. Lee and Grimson [5] introduce a new feature representation method in gait recognition. It divides human silhouette image into 7 parts, which are fitting with 7 ovals. Each oval is represented with its 2 centroid coordination, the ratio of its long and short axis, and the direction of long axis. Wang Liang [6] employs high dimension eigenvector of gait, which is consisted from centroid and edge of silhouette image. And the gait recognition is reduced by dimensionality of high dimension eigenvector with principle component analysis and similarity measurement of low dimension vector.

Wang Liang acquires satisfied result of recognition with principle component analysis of high dimensional feature space. However principle component analysis, only uses several features are used to express original gait information, and different modes dose not have optimal separability after reducing dimensionality. Therefore, this paper introduces gait recognition algorithm with PCA and LDA based Wang Liang's research. The PCA is optimal for gait feature, and LDA is best for modes optimal classification. e.

II. GAIT FEATURE EXTRACTION

A. Detection of Moving Objects

Due to the background is static during taking video, this paper employs background subtraction algorithm to measure pedestrian gait. A completed gait detecting process is demonstrated in Figure 1.

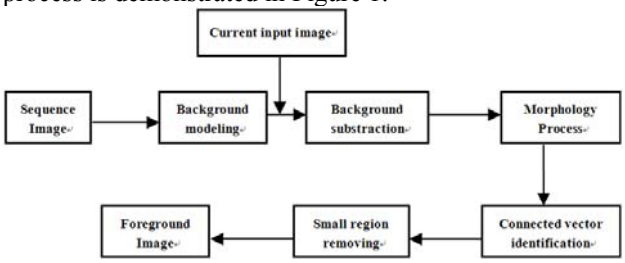


Figure1 Gait Detection Process

The median algorithm is adopted in the extraction of background image from sequence image, background image is dynamically generated from image sequence. The difference between foreground and background images is used to detect the moving object. Inorder to eliminated non-background pixels and noise that caused by existence of silhouette and other reasons at the static ackground, to filter the noise and background substraction. we apply noise interference the morphological operator to filling the holes. Then by operating the methed of 8-connected region contour tracing, a simply connected moving object is extracted.

B. Gait silhouette expression

The silhouette shape varing with time reflects the in struct gait feature of pedestrian. To reduce the complexity of calculation, it transforms 2-dimensional change of silhouette shape into 1-dimensional distance signal, which is used to approximately express the spatiotemporal mode of gait movement. Firstly, each point on the boundary of silhouette is regarded as a vector and showed as equation 1 in plural form (assuming there are N points in the silhouette):

Foundation Project::
National Nature Science Foundation(No.60702079);
Hubei Province Education department financing projects(D20081407)

$$z_i = x_i + jy_i \quad i = 1,2,\cdots,N \tag{1}$$

(1) The centroid. he centroid of silhouette is calculated as follows:

$$x_c = \frac{1}{N}\sum_{i=1}^{N} x_i \quad , \quad y_c = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{2}$$

Where, （xc , yc）is the coordination of centroid, N is the pixels number of silhouette, and （xi , yi）is the point coordination of silhouette.

(2) Distance. Taking the vertex as the stating point, calculate distance between each silhouette point and centroid point in the counterclockwise direction.

$$d_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \tag{3}$$

(3) Normalization. In order to eliminate image size difference, the distance signal is normalized by D1, which any is distance between vertex point and centroid point being unit 1. It is given by:

$$D_i = \frac{d_i}{d_1} \quad , \quad i = 2,\cdots N \tag{4}$$

As a result of silhouette shape and the number of silhouette varying with time, the methed of equal interval sampling is used to make sure very gait sequence has identical distance signal number.

C. Gait feature analysis

As a kind of spatiotemporal movement, gait sequence could be considered as a mode, constructed by a set of static postures, which contains large amount of spatiotemporal information. Gait feature of movement is a high dimension eigenvector, which not only need large storage room, but also increases the space and time complexity of system. Therefore, to reduce the dimensionality of gait feature, some process is necessary carry out before recognition. Currently, feature space transformation technique is being widely used in human face recognition and gait analysis. This paper employs PCA and LDA methods to perform training and projecting on original gait feature. Firstly, it reduces dimensionality of high dimensional gait feature with PCA, and then performs optimal classification on low dimensional space with LDA algorithm.

(1) PCA training

The number of normalised gait category is named as c, and each category represents a distance signal sequence of some person's gait mode. Let Di,j indicate the jth distance signal in category i, each distance signal contain M normalized pixel points (dimensionality), the sample number of distance signal in category i is Ni, then the total training sample number is

$$N = N_1 + N_2 + \cdots + N_c .$$

The whole training set is

$$\boldsymbol{D} = \left\lfloor \boldsymbol{D}_{1,1}, \boldsymbol{D}_{1,2}, \cdots, \boldsymbol{D}_{1,N_1}, \boldsymbol{D}_{2,1}, \cdots, \boldsymbol{D}_{c,N_c} \right\rfloor ,$$

$$\boldsymbol{D} \in R^{M \times N} ,$$

in which the mean value—$\boldsymbol{m}_{\mathrm{D}}$ and covariance matrix—Σ are

$$\boldsymbol{m}_{\mathrm{D}} = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N_i} \boldsymbol{D}_{i,j} \tag{5}$$

$$\Sigma = \frac{1}{N-1}\sum_{i=1}^{c}\sum_{j=1}^{N_i} (\boldsymbol{D}_{i,j} - \boldsymbol{m}_{\mathrm{D}})(\boldsymbol{D}_{i,j} - \boldsymbol{m}_{\mathrm{D}})^{\mathrm{T}} \tag{6}$$

In above, Σ is real symmetric matrix, and $\Sigma \in \mathrm{R}^{M \times M}$. According to singular value decomposition (SCD) theory, the nonzero eigenvalue of M are $\lambda_1, \lambda_2, \cdots, \lambda_M$ and its corresponding eigenvector are $\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_M$. The former eigenvectors with relatively large eigenvalue have relatively large change in training mode. And higher order eigenvector is to indicate relatively little change. Considering with storage and effectivity of calculation, this paper puts threshold value T_s on accumulated variance curve to remove those relatively small eigenvalue and its corresponding eigenvector.

If we select k (k<M) biggest eigenvalues and their eigenvector, the eigentransformation matrix $\boldsymbol{W}_{\mathrm{PCA}}$ can be constructed as,

$$\boldsymbol{W}_{\mathrm{PCA}} = \left[\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_k\right] \tag{7}$$

By project each original distance signal $\boldsymbol{D}_{i,j}$ into k-dimensional eigenspace, we acquire $\boldsymbol{R}_{i,j}$.

$$\boldsymbol{R}_{i,j} = \boldsymbol{W}_{\mathrm{PCA}}^{\mathrm{T}} \cdot \boldsymbol{D}_{i,j} = \left[\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_k\right]^{\mathrm{T}} \boldsymbol{D}_{i,j} \tag{8}$$

(2) LDA training

After training and projecting D by PCA, a new sample collection, $\boldsymbol{R} = \left\lfloor \boldsymbol{R}_{1,1}, \boldsymbol{R}_{1,2}, \cdots, \boldsymbol{R}_{1,N_1}, \boldsymbol{R}_{2,1}, \cdots, \boldsymbol{R}_{c,N_c} \right\rfloor$ $\boldsymbol{R} \in \mathrm{R}^{k \times N}$ could be obtained, and its dimensionality is reduced from M to k. Based on equation (9) and (10), new sample's $\boldsymbol{S}_{\mathrm{w}}$ —within-class scatter matrix and $\boldsymbol{S}_{\mathrm{b}}$ —inter-class scatter matrix are being calculated as following,

$$\boldsymbol{S}_{\mathrm{w}} = \sum_{i=1}^{c}\sum_{j=1}^{N_i} p_i (\boldsymbol{R}_{i,j} - \boldsymbol{m}_i)(\boldsymbol{R}_{i,j} - \boldsymbol{m}_i)^{\mathrm{T}} \tag{9}$$

$$\boldsymbol{S}_{\mathrm{b}} = \sum_{i=1}^{c} p_i (\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^{\mathrm{T}} \tag{10}$$

Then the matrix $\boldsymbol{S}_{\mathrm{w}}^{-1}\boldsymbol{S}_{\mathrm{b}}$ is obtained. According to the theory of singular value decomposition, we could calculate P (P<k) non-zero eigenvalue $\lambda_1, \lambda_2, \cdots, \lambda_P$ and their corresponding eigenvector $\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_P$. Some relatively large eigenvalue and with their eigenvector have comparatively separability, and higher order

eigenvector has comparatively lower separability. By setting threshold value T_f , we select t maximum eigenvalue in front with their eigenvector to construct eigentransformation matrix $\boldsymbol{W}_{\mathrm{LDA}}$, which is

$$\boldsymbol{W}_{\mathrm{LDA}} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_t] \quad (11)$$

By projecting each new sample— $\boldsymbol{R}_{i,j}$ to t-dimensional eigenspace, we obtain the projected signal $\boldsymbol{P}_{i,j}$:

$$\boldsymbol{P}_{i,j} = \boldsymbol{W}_{\mathrm{LDA}}^{\mathrm{T}} \cdot \boldsymbol{R}_{i,j} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_k]^{\mathrm{T}} \boldsymbol{R}_{i,j} \quad (12)$$

Put $\boldsymbol{R}_{i,j} = \boldsymbol{W}_{\mathrm{PCA}}^{\mathrm{T}} \cdot \boldsymbol{D}_{i,j}$ into equation 12, and it gets

$$\boldsymbol{P}_{i,j} = \boldsymbol{W}_{\mathrm{LDA}}^{\mathrm{T}} \cdot \boldsymbol{W}_{\mathrm{PCA}}^{\mathrm{T}} \cdot \boldsymbol{D}_{i,j} \quad (13)$$

$\boldsymbol{P}_{i,j}$ is a point in the t-dimensional eigenvector、each gait sequence appears as a track in eigenspace. It is obvions that PCA training enormously reduces the dimensionality of sample and optimal classification after LDA training.

III. GAIT RECOGNITION

A. Similarity measurement

Due to Gait being spatiotemporal movement, the structure and time-shifting characteristic of gait are used to captured by STC (Spatio-temporal Correlation). Two random gait sequences $\mathrm{D}_1(t)$ and $\mathrm{D}_2(t)$, their projected tracks in eigenspace of $\boldsymbol{W}_{\mathrm{PCA}} \cdot \boldsymbol{W}_{\mathrm{LDA}}$ are $\mathrm{P}_1(t)$ and $\mathrm{P}_2(t)$ separately, and the their similarity measurement could be defined as [8]:

$$d^2 = \min_{a,b} \sum_{t=1}^{T} \| \mathrm{P}_1(t) - \mathrm{P}_2(at+b) \|^2 \quad (14)$$

In equation 14, $\mathrm{P}_2(at+b)$ is vector track which originates from the expanding, shrinking and displacement of $\mathrm{P}_2(t)$ on time. Parameters a and b are depended on change of velocity and phase position among different sequences.

Furthermore, this paper applies NED (Normalized Euclidean Distance) between the projected centroid to measure the similarity of different sequences. For every gait sequence, projected centroid C_i is acquired by the projection of every single frame in average sequence.

$$C_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \boldsymbol{P}_{i,j} \quad (15)$$

With two projected track of sequences— $\mathrm{P}_1(t)$ and $\mathrm{P}_2(t)$, and equation 15, the corresponding projected centroid— C_1 and C_2 could be worked out. Every projected centroid indirectly indicates a main silhouette shape of such category, and it demonstrates constructing model of gait mode. The definition of NED is:

$$d^2 = \| \frac{C_1}{\| C_1 \|} - \frac{C_2}{\| C_2 \|} \|^2 \quad (16)$$

B. Classification method

Among mode classification methods, nearest neighbor (NN) and K-nearest neighbor (KNN) are classical methods. NN is calculating Euclidean distance between unknown sample and training sample, and classifies the unknown sample to the category that has the minimum Euclidean distance with unknown sample. KNN is calculating k neighbors around the unknown sample, and classifies the unknown sample to the category that majority neighbors belongs to. In the experiment, NN and KNN are separately applied in the classification and identification of former used STC and NED.

IV. EXPERIMENT RESULT

A. Experiment data

The gait database in this paper contains 30 persons and 4 sequences for each visual angle of every person, totally 360 (30×4×3) sequences. These true color images are shot at 25fps, and original size is 320×240. Every sequence length depends on the duration time of appearance in the visual field, and its average time is around 80 frames. During the experiment, we selects 120 sequences of silhouette, and all video sequences has been transformed to grey scale image by Matlab.

B. Experiment result and analysis

For each gait sequence, gait measurement is used to monitor person's movement. Before the training and projection, we transform the 2-dimensional silhouette image sequence into 1-dimensional distance signal sequence. For the 120 sequence of silhouette in the experiment, we take one sample sequence as testing sample, and train the rest—119 sequence. After training and projecting, the testing sample is classified by its similarity against training samples this process repeats 120 times. This paper adopts the FERET evaluation agreement to report the recognition results. The statistical property of performance is reported as iterative matching values. In following figures, horizontal axis is order k, and vertical axis is iterative matching value of correct match. Figure2 illustrates the recognition result between PCA training and PCA plus LDA training. Figure3 shows the recognition results between this paper's algorithm and Wang Liang's.

In fgiure2, the recognition ratio obtained by the PCA methed is much worse than that based on the methed of PCA and LDA. This paper performs PCA training and projecting for the original gait samples, without LDC training. PCA reduce dimensionality of original data. The projected data could basically represent the original gait information, but its distribution in eigenspace is not compact. LDA could raise the veracity of recognition

with its better compaction on identical gait mode and larger separation on different gait modes after projecting.

In figure3, if we do the similarity measurement by STC with the NN classification, the discrimination of Wang Liang's algorithm is slightly lower than that of this proposed algorithm, but the discrimination of this paper's algorithm is much higher than that of Wang Liang's method. Considering the sample of this paper containing 120 sequence, which are more than that in Wang Liang's experiment, the proposed algorithm has leading advantage in LDA training that makes each gait mode more separable.

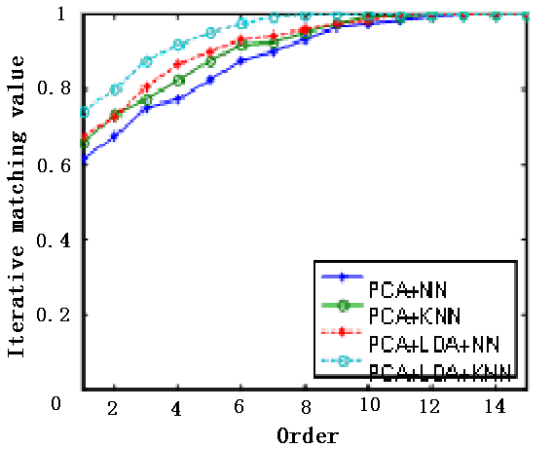


Figure2. recognition result with different training method

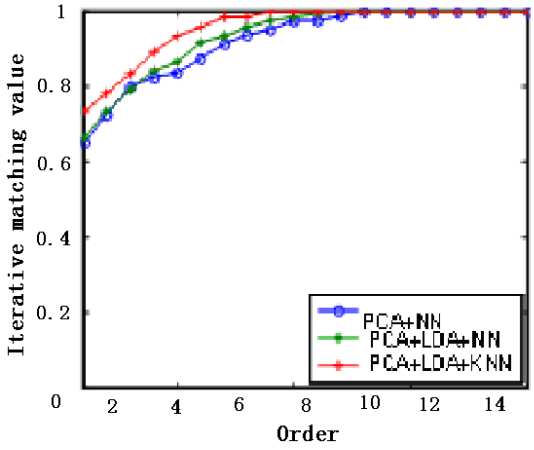


Figure3. Recognition result on different algorithm

V. CONCLUSION

This paper introduces a new gait recognition algorithm based on PCA and LDA. The result of experiment demonstrates that the gait recognition algorithm of PCA and LDA are better than that of PCA. After the analysis and comparison of experiment result, this paper's algorithm has achieved satisfied recognition result.

REFERENCES

[1] Dittrich W H. Action categories and the perception of biological motion[J]. Perception, 1993,22:15-22.

[2] Niyogi S A, Adelson E. H. Analyzing and recognizing walking figures in XYT[C]. Proceeding of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition,Seattle, 1994, pp:469-474.

[3] Kale A, Rajagopalan A, Cuntoor N, Krger V. Gait-basedrecognition of humans using continuous HMMs[C]. In Proc. Of ICAFGR2002, Washington D C, USA, 2002, pp:336-341.

[4] C. BenAbdelkade, R. Culter, H. Nanda, L. Davis. EigenGait: motion-base recognition of people using image self-similarity[C]. Proc. of AVBPA2001, Halmstad, Sweden, June 2001, pp:284-294.

[5] Lee L, Grimson W E L. Gait Appearance for Recognition[C]. ECCV Workshop on Biometric Authentication, Copenhagen, Denmark, 2002, pp:143-154.

[6] WangLiang, HuWeiming,TanTienv Gait-Based Human Identificati0n[J]. Chinese Journal of Computer，2003，26（3）：354-360

[7] Kuno Y, Watanabe T, Shimosakoda Y, Nakagawa S. Automated detection of human for visual surveillance system[C]. Proc. of Intl. Conf. on Pattern Recognition, 1996, pp:865-869.

[8] BianZhaoqi, ZhangXuegong. Pattern recognition [M].BeiJin: Tsinghua University publishing house,1999.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 128-133

Make Palm Print Matching Mobile

Fang Li[1] , Maylor K.H. Leung[2] , and Cheng Shao Chian[3]
[1] School of computer engineering , Nanyang Technological University , Singapore
asfli@ntu.edu.sg
[2] School of computer engineering , Nanyang Technological University , Singapore
asmkleung@ntu.edu.sg
[3] School of computer engineering , Nanyang Technological University , Singapore
Y060043@ntu.edu.sg

***Abstract*—With the growing importance of personal identification and authentication in today's highly advanced world where most business and personal tasks are being replaced by electronic means, the need for a technology that is able to uniquely identify an individual and has high fraud resistance saw the rise of biometric technologies. Making biometric-based solution mobile is a promising trend. A new set of palm print image database captured using embedded cameras in mobile phone was created to test various segmentation techniques on their robustness. The improved square-based palm print segmentation method was successfully implemented and integrated into the current application suite. Comparing to the two segmentation methods that are based on boundary tracking of the overall hand shape that has limitation of being unable to process palm print images that has one or more fingers closed, the system can now effectively handle the segmentation of palm print images with varying finger positioning. The high flexibility makes palm print matching mobile to be possible.**

***Index Terms*—Palm print; segmentation; mobility;**

I. INTRODUCTION

Personal identification and authentication have become a common task in today's highly advanced world where more and more day-to-day personal and business activities have been computerized. Traditional identification and authentication systems relies on either a token item (For e.g. a security pass card) or some knowledge only the user would know (For e.g. passwords). Such systems are usually expensive in terms of time and resources to maintain and expand its usage. The most critical flaw of these systems is that since they do not use any inherent characteristics or attributes of the individual user, they are unable to differentiate between an authorized personnel and an impostor who have fraudulently come to possess the token or knowledge (Such as stolen credit card or lost password). As such, these problems have led to system developers and researchers to explore into alternative solutions, and thus the intensified research on biometric identification and authentication systems.

Following this initial foray into biometric research, several forms of biometric systems based on different physiological or behavioral characteristics have been developed. The first commercial system, Identimat was developed in the 1970s [7]. The system was based on the measurement of the shape of the hand and the lengths of the fingers as the basis for personal identification. Following that, various forms of biometric systems such as fingerprint-based systems and iris, retina, face, palm print, voice, handwriting and DNA technologies joined in over the years.

Among the leading biometric technologies, fingerprint-based system is the most prominent and widely used biometric technology, encompassing a market share of 58% in 2007 (A combine percentage of fingerprint and AFIS/Livescan technologies) [1]. The small size of the fingerprint-based device, ease of use and high accuracy has made it largely popular; however, as with most biometric solutions, there are certain drawbacks to it. It is commonly found in most people that a layer of oil secretion or perspiration which emits from microscopic pores residing on the tiny ridges of the fingers will cover the surface of the fingerprint areas. As the resolution required for the fingerprint images are relatively high at approximately 500 dpi [7], this layer of secretion will render the fingerprint image capturing device useless or less effective in most cases. There are also cases whereby fingerprints wear away due to work or fraudulently scarred, all these will lower the effectiveness of fingerprint based systems.

In this project, we explore a relatively new biometric technology that employs palm print as the physiological characteristic that is used to differentiate between each unique individual. Palm prints are rich in features such as principal lines, wrinkles, ridge, datum points and minutiae points, all of which could be extracted at relative low resolution. Palm prints also have a much larger surface area as compared to fingerprints, which indicates that more features could be extracted from it, adding higher level of accuracy to it. These advantages place palm print-based technology as a promising biometric identification system.

Palm print recognition is an effective biometric technology that is gaining widespread acceptance and interest from researchers all over the world. As with most other biometric technologies, the process of palm print identification includes various stages from data acquisition, data pre-processing, feature extraction to matching process.

The main aim of this research is to improve the segmentation process to increase the system robustness. By implementing and integrating a new square-based palm print segmentation method into the previous application suite, the system is now able to overcome the

limiting problem of failure to process palm print images with closed fingers, thus increasing the flexibility of the system and in turn open up the possibility of bringing the palm print technology mobile. A new set of palm print image database captured using embedded cameras in mobile phone was created to find most robust segmentation technique to let palm print matching mobile come true.

II. Overview of Palm Print Authentication System

Palm print recognition is an effective biometric technology that is gaining widespread acceptance and interest from researchers all over the world. As with most other biometric technologies, the process of palm print identification includes various stages from data acquisition, data pre-processing, feature extraction to matching process. The system overview is shown in Figure 1.

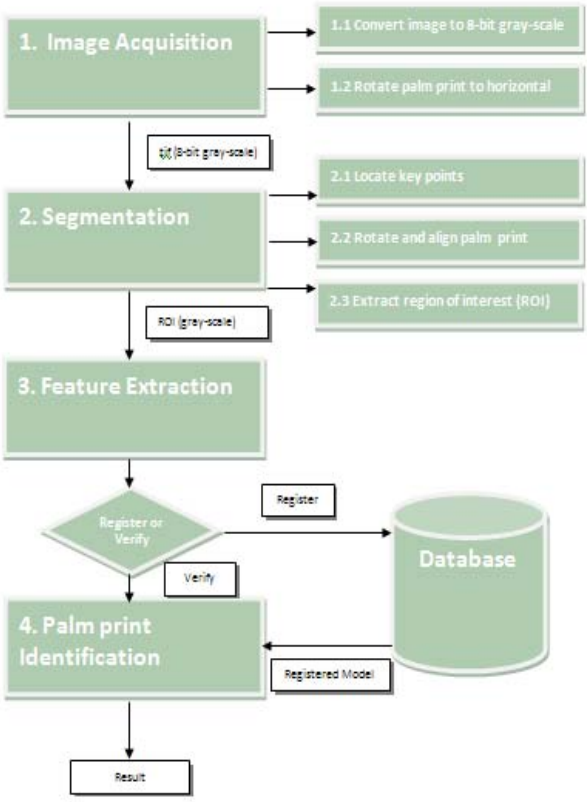


Figure1: Palm print authentication system overview

III. Pre-processing: Segmentation

Palm prints may show certain degree of distortion as the image may be captured at different times and rotated at different angles, furthermore, it could also be affected by varying conditions in terms of temperature, humidity and lighting condition, as such, even an image of the same palm, we could end up with a totally different looking image altogether.

Palm print preprocessing, the segmentation process, involves the correction of such distortion and placing all the palm prints in the database under the same coordinate system and orientation such that the proper expected area of each palm print can be extracted for use in accurate feature extraction and matching, greatly improving the efficiency and correctness of the identification system.

Palm print data can be broadly classified into two categories, offline and online, and the image quality of inked offline palm prints[4] is different from that of an online palm prints, thus different segmentation approach is required for each type of images. Inked palm print is not a good choice for mobile matching, so we only discuss online palm print in this research.

Online palm print segmentation methods can be further classified into two different classes: square-based segmentation and inscribed circle-based segmentation [7]. As the circle-based approach consumes a significantly higher amount of computation resources, based on this experimental outset, we will only focus on square-based segmentation approach throughout this research.

The basic idea of square-based segmentation technique is to determine key gaps-between-fingers point on a palm print, thereafter an orthogonal coordinate system is set up by using these key points. Finally, a square with a fixed size, known as the region of interest (ROI) or central part sub-image of the palm print is extracted under this coordinate system. All the pixels within this ROI are retained for further processing whereas the area outside the window are ignored and discarded. The essential rule in this extraction process is that the portion of the image extracted should be available in all palm prints from the database and there are sufficient palm print features for extraction and comparison.

Under the implementation of the square-based segmentation in the device-constraint system, which is used for performing palm print segmentation against the PolyU-ONLINE-Palm print-II research benchmark database [5], the steps can be summarized as follows:

Step 1: Extract boundary of palm print

Given an 8-bit gray scale palm print image as input, as shown in Figure 2(a), the image is converted to a binary image map according to the computed threshold value. The image is then passed through a filter and morphological thinning to produce the extracted edge image, Figure 2(b).

Step 2: Get pairs

Pairs of lines are identified and eventually the baseline of the palm print is located.

Step 3: Get key points

The key points are identified according to the positioning of the two end points of the baseline from the palm print image.

Step 4: Construct area

Construct a line through the two key points identified through the earlier step to retrieve the Y-axis of the coordinate system, and then a second line is drawn across the midpoint of the two key points which is perpendicular to the Y-axis, to determine origin of the coordinate system. A fixed size square is then determined according to this coordinate system for clipping.

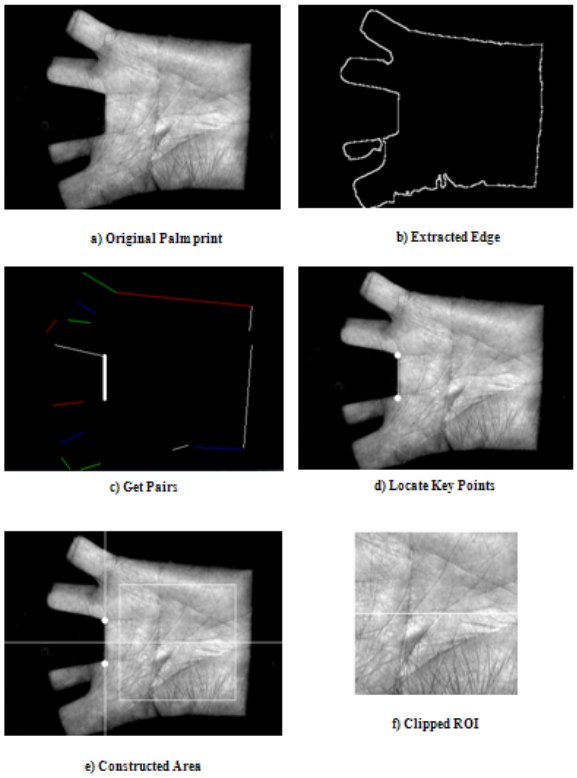


Figure2: Major steps of device-constraint segmentation method

There are several other square-based segmentation approaches, which relies on the definition of three key points [6], which are the three mid points between the four fingers excluding the thumb. In general, all the approaches involves finding the key points using boundary tracking algorithm and lining up two key points to form the y-axis, thereafter a second line, which is the x-axis is drawn perpendicular to the y-axis through the middle point to form the origin. Subsequently a fixed size sub-image is extracted from the derived coordinate system.

All segmentation techniques principally rely on the determination of the key gaps-between-fingers to draw up the coordinate system through the use of boundary tracking algorithm. Thus this imposes a limitation in that for a palm print image to be properly segmented, the fingers in the image need to be sufficiently separated in order for the boundary tracing to work, and determine the key points accurately.

Such requirement led to the development of image acquisition devices that utilizes pegs [7] to restrict the movement and positioning of the hand during the acquisition process in order to improve the image quality, and to ensure that the fingers are properly separated. However, such devices are normally fixated to a site, and too bulky to move around for usage. This lack of mobility consequently results in the restrictive applications of palm print authentication system.

IV. IMPROVED SQUARE_BASED PALM BASED PALM PRINT SEGMENTATION METHOD SEGMENTATION

The aim of this research is to implement an improved segmentation technique that is robust enough to overcome this reliance of a standard image acquisition device, and thus able to make use of the ubiquitous digital cameras and embedded cameras in mobile phones to perform the image capturing process, which in turn, will widen the scope of applications for palm print-based systems.

The improved square-based palm print segmentation method is first described in [8]. The algorithm first determine a local area in which the boundaries of fingers section, then the palm print image is aligned by using the outside boundary of the palm as a reference line. Lastly, a square sub-image of the palm print image is extracted based on a percentage of the size of the palm. This approach eliminates the problem of palm print images not being able to be properly segmented (extract region of interest) when one or more fingers are closed together.

The improved square-based segmentation technique consists of the following steps:

Step 1: Gray image to binary image.

This is a standard step as shown in Figure 3 and 4.

Figure3: Original Palmprint Image

Figure4: Binarized Image Map

Step 2: Extract a local area

As described in [8] and shown in Figure5 and 6, local areas are extracted.

Step 3: Detect three key points

Based on the extracted boundary of the local area, locate the three key points K1, K2 and K3 as shown in Figure 7.

Step 4: Align palm print image

Using two key points nearer to the outside boundary of the palm, the position of the start point of the reference line, P1 is determined. Point P3 can be determined after that. It is located at the other side of the palm, and on the same row as P1. The length of the line P1P3, β, is the approximate width of the palm. The point P2 is set by tracing the outside boundary of the palm from start point P1. The line P1P2 is two-third of β.

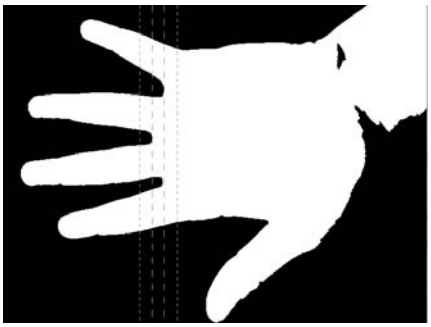


Figure5: Finding of Smallest Local Area and Local Area

Figure6: Extracted Local Area

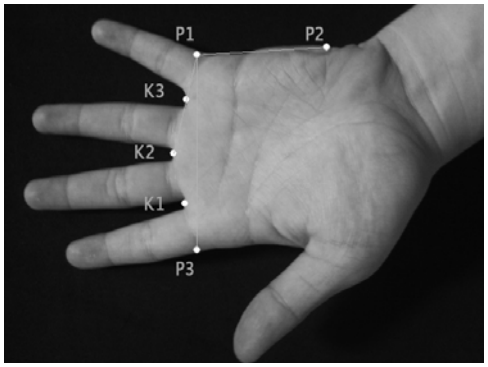


Figure7: Plotted Points

The direction of the line P1P2 is taken as the direction of the palm print and the palm print image is rotated so that all palm prints face the same direction. P1' is the adjusted reference point in the aligned image. P4 is the point located in the other side boundary and on the same row with P1' in the aligned image. The length of P1'P4 is taken as the real width of the palm.

Step 5: Construct coordinate system and extract central part sub-image

Line up P1' and P4 to obtain the x-axis of the palm print coordinate system, the center point of P1' and P4 is determined as the origin of the coordinate system. As shown in Figure 8, a perpendicular line from the x-axis through the origin will be the y-axis. After this coordinate system is set, the square sub-image of the central part which is seventy percent of the width of the palm is extracted.

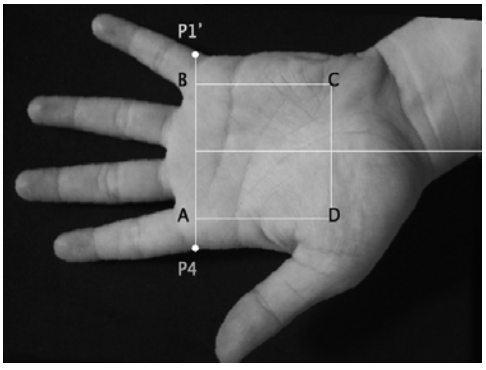


Figure8: Coordinate System Setup

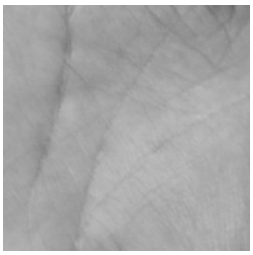


Figure9: Extracted region of Interest

V. EXPERIMENT AND RESULTS

As the main aim of this research is to implement segmentation method that is capable of handling palm print images that are varying in terms of fingers positioning. A new Mobile Palm print Database with one thousand five hundreds photos is formed.

In this research, palm print images are captured using three mobile embedded cameras with different resolutions from two different mobile phones. During the image capturing process, no fixed pegs were used to restrict the movement, rotation and stretching of the hands. Each device is used to capture images of both hands from thirty subjects. Figure 10, 11, and 12 show the sample images captured by different mobile phones. With each hand, five photos are taken for each of the five different positioning of the hands in order to test the robustness of the algorithm. Some samples of various finger positions are shown in Figure 13.

Figure10: Palm print captured using D810 VGA camera

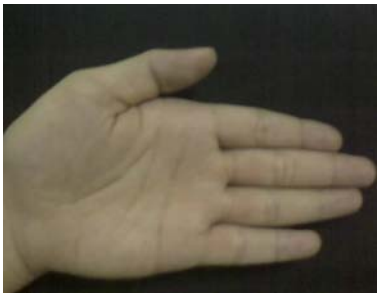


Figure11: Palm print captured using D810 2MP camera

Figure 12: Palm print captured using SGH-i900 5MP camera

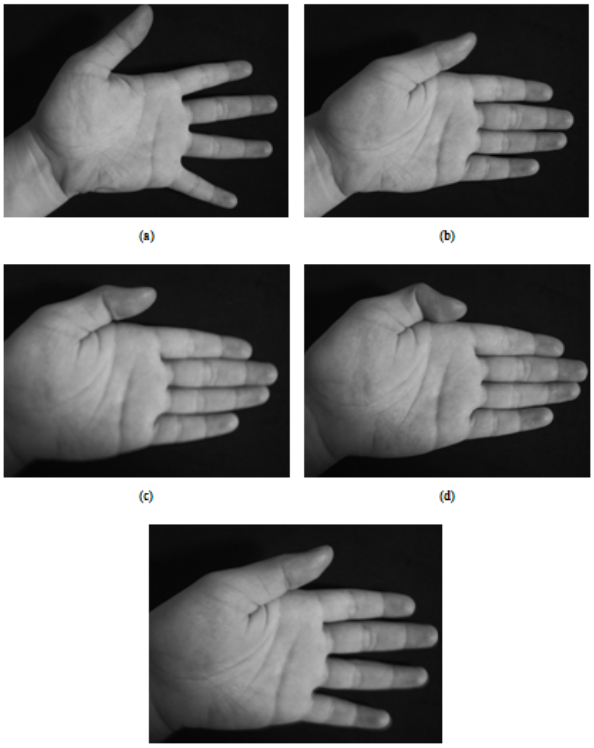


Figure 13: Various positioning of fingers. (a) Fingers opened widely, (b) Fingers opened slightly, (c) Fingers closed slightly, (d) Fingers closed tightly and (e) Fingers opened naturally

As shown from Figure 14 to 15, to test the improved segmentation method, an image from the Mobile Palm print Database that has finger regions tightly closed together was used. The steps in Figure 16 show that the improved scheme can extract the region of Interest successfully.

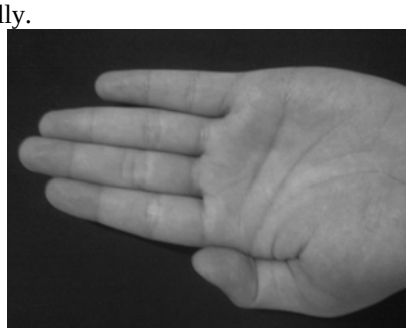


Figure 14: Input palm print image to test the two segmentation methods

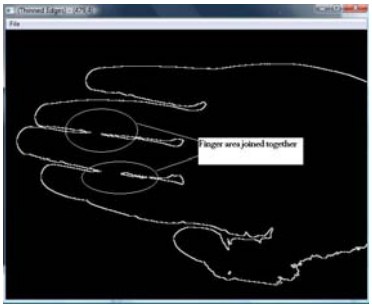


Figure 15: Thinned edges map shows fingers are closed at certain area

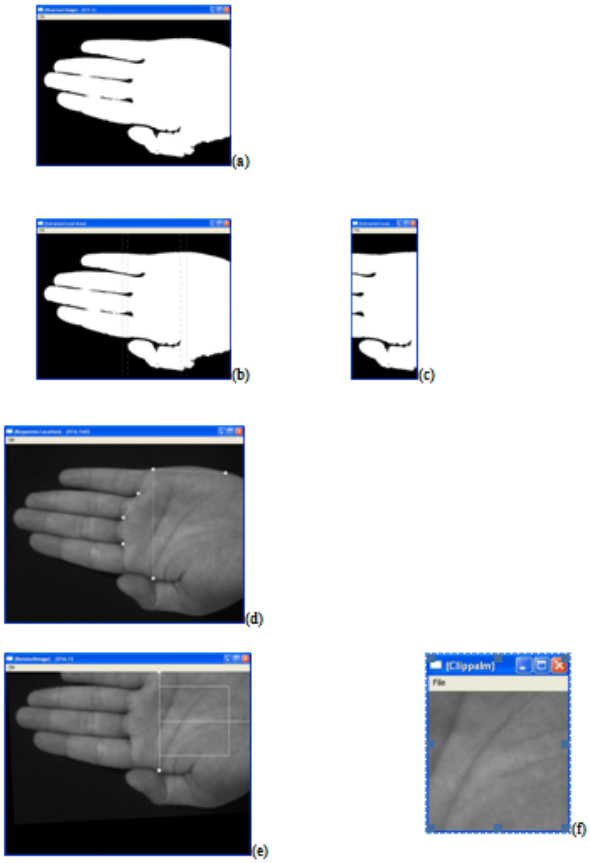


Figure 16: Improved segmentation method. (a) Binarized palm print image, (b) Smallest effective local area and local area identified, (c) Local area extracted, (d) Various key points calculated and plotted, (e) Palm print aligned according to coordinate system and (f) Extracted region of interest

VI. Conclusion

The improved square-based palm print segmentation method was successfully implemented and integrated into the current application suite. In comparison to the previous two segmentation methods that are based on boundary tracking of the overall hand shape that has limitation of being unable to process palm print images that has one or more fingers closed, the system can now effectively handle the segmentation of palm print images with varying finger positioning. It opens up the possibility of bring the palm print technology mobile.

Through the experiments and findings in this research, it was found that the images captured with latest model of the mobile camera at five mega pixels, the verification rate

was close and comparable to those of the palm print images in the research benchmark. Thus, with the certain continuous advancement in mobile camera technologies, it would be in no time that the usage of palm print authentication in mobile systems be proved as viable and practical for widespread applications.

REFERENCES

[1] Biometric technologies - an introduction. Intelligence, Acuity Market. 2007, Biometric Technology Today, p. 9.

[2] Adams Kong, David Zhang and Guangming Lu. A Study of Identical Twins' Palm prints. s.l. : Springer-Verlag Berlin Heidelberg, 2005.

[3] Human identification in information systems: Management challenges and public policy issues. Clarke, R. 1994.

[4] W.X. Li, D. Zhang and Z.Q. Xu. Image alignment based on invariant features for palm print identification. Signal Processing: Image Communication. 2003, Vol. 18,

[5] H. K. Polytechnic University,. Palm print database, 2005. Biometric Research Center Website. [Online] 2005. [Cited: November 25, 2008.] http://www4.comp.polyu.edu.hk/~biometrics/.

[6] Li, W.X. Authenticating Personal Identites Using Palm Print Recognition. s.l. : Hong Kong Polytechnic University, 2004.

[7] Zhang, David D. Palmprint Data. Palmprint Authentication. s.l. : Springer Science + Business Media Inc., 2004.

[8] An Improved Square-based Palmprint Segmentation Method. Yanxia Wang, Qiuqi Ruan and Xin Pan. Xiamen, China : IEEE, 2007.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 134-138

Agent Based Distributed Intrusion Detection System (ABDIDS)

Yu Lasheng , and MUTIMUKWE Chantal
Central South University (CSU), Department of Computer Science, Changsha, 410083, China
ley462@163.com, cmutimukwe@gmail.com

***Abstract*—This paper introduce (ABDIDS), a simple pattern attack ontology that allows agent based intrusion detection system to detect network traffic anomalies at a higher level more than most current intrusion detection systems do. The cooperative agent architecture has been presented. It has been shown how some attributes in network communication can be used to detect attacks. Finally, the benefits of using the proposed values in attack pattern Ontology within intrusion detection system have been illustrated**.

***Index Terms*—intrusion detection, agents, network attacks, ontology.**

I. INTRODUCTION

As the use of computer system increases; intrusions (worms attack, Denial of Services, port scans, etc） against them increase too. Intrusion is a set of actions which attempt to compromise the confidentiality, integrity or availability of a resource [1]. That is the reason of many demands of effective and powerful intrusion detection system.

During the last two decades, many strategies and methods have been developed [2]. First research on computer-aided intrusion detection goes back to the 80's [1].

However, using current generation of IDS are continuously overwhelmed with a vast amount of log information and bombarded with countless alerts. The capacity to tolerate false positives and correctly respond to the output of current IDS is debatable [2]. There are those who even postulate that traditional IDS not only have failed to provide an additional layer of security but have also added complexity to the security management task.

Therefore, there is a compelling need for developing a new generation of tools that help to automate security management tasks such as the interpretation and correct diagnosis of IDSes output.

This approach proposes ABDIDS as a fully distributed system made by set of nodes with three types of agents: Monitoring Registry Agents (MoRA), Monitoring Agents (MoA) and managing agents (MA). It is commonly known that in the case of worm attack there occur at least two kinds of anomalies: in observed traffic characteristics and in communication scheme which tends to be constant under normal conditions. In this work, the attack recognition is being made on the basis of them.

The MoA agent's algorithm for decision making process is invoked periodically and uses observed values as input data. MoA also stores acquired values thus creating the history of system behavior.

II. RELATED WORKS

Crosbie and Spafford were the first to propose autonomous agents in the context of intrusion detection. Their initial proposal evolved to become AAFID [4]. Other works such as Cooperating Security Managers have proposed a multi-agent system to handle intrusions instead of only detecting them [3].

However, in these works agents lack reasoning capabilities and are used for mere monitoring. More sophisticated agents with richer functionality were introduced by [12]. Different taxonomies of computer security incidents have been proposed [13].

An ontology centered on computer attacks was introduced in [12].That ontology provides a hierarchy of notions specifying a set of harmful actions in different levels of granularity from high level intentions to low level actions.

III. ABDIDS: AGENT BASED DISTRIBUTED INTRUSION DETECTION SYSTEM

A. Intrusion Detection System

This paper proposes an Agent based distributed Intrusion detection system as an important component of defensive measures protecting computer systems and networks from abuse.

This work describes the security assessment of a network system requires application of complex and flexible mechanisms for monitoring values of system attributes that have an influence on the security level of all network system.

Another important element is an effective computational mechanism for evaluating the states of system security on the basis of incomplete, uncertain and inconsistent resources has been considered. Finally, the algorithms of machine learning to detect new intrusions pattern scenarios and recognize new symptoms of security system breach in order to update the security system knowledge base must be defined.

B. Agent Based System

One of the aims of this paper is to propose a framework for agent based distributed Intrusion Detection System. An agent is an autonomous, collaborative, software entity. They are designed to allow

software systems to delegate tasks and undertake roles in an intelligent manner [5).

Agent is considered as proactive, because it does not simply act in response to its environment but is able to exhibit goal-directed behavior by taking initiative. Moreover, if necessary an agent can be mobile, with the ability to travel between different nodes in a computer network. It can be truthful, providing the certainty that it will not deliberately communicate false information. It can be benevolent; always trying to perform what is asked of it. It can be rational, always acting in order to achieve its goals and never to prevent its goals being achieved, and it can learn, adapting itself to fit its environment and to the desires of its users [6].

In this paper it is assumed that the network system is consisted of the set of nodes. There are also three types of agents in our agent based system: Monitoring Registry agents (MRoA), monitoring agent (MoA) and managing agent (MA). See Fig1.

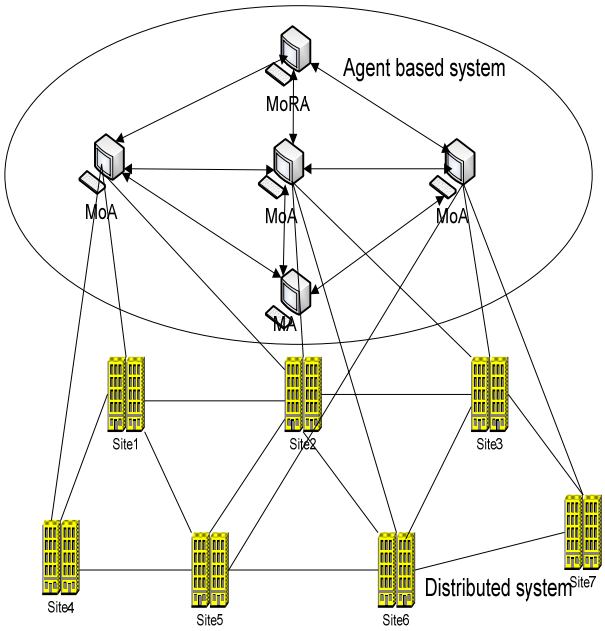


Figure1: ABDIDS Architecture.

MoRA is responsible to initialize and identify the Monitoring Agents. It defines the current status of each monitoring agent and maintains information about them (eg: MoA-Id, location). It is required that all monitoring agents must register with the MoRA in order to get an Identification number (MoA-Id). The Monitoring Agent status is based on two parameters : *alive* and *reachable.* See Fig2.

Monitoring agent's purpose is to collect data on security related events on different nodes and transmit these to the Managing Agent. Each MoA works within his own area of responsibility. See Fig 2.

Managing agents receive reports from MoA and processes and correlates these reports to detect intrusion. See Fig 2

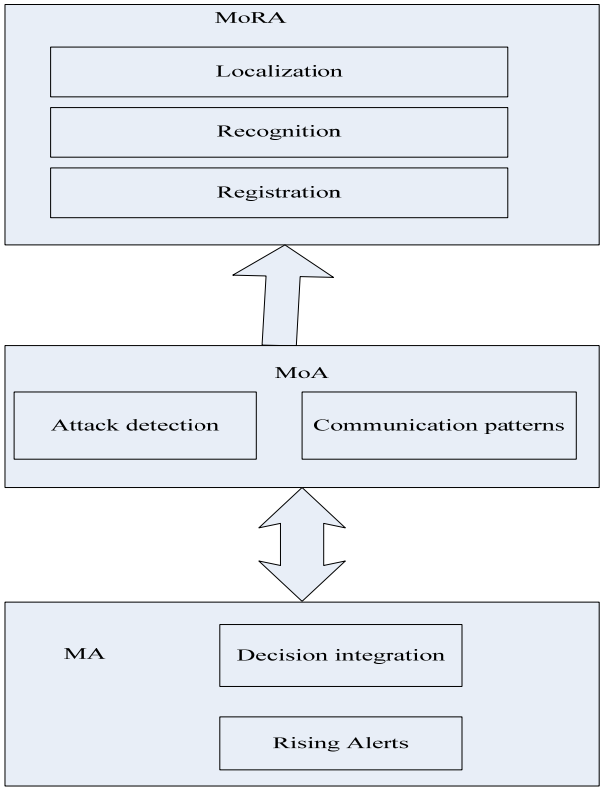


Figure2: the functional structure of Agent

C. Network Attack

The attributes which have been selected because they are especially important (because their rapid change during typical attacks) and used during process of anomaly detection are [7]:

- source and destination IP address,
- source and destination port,
- number of bytes and packets sent to the remote hosts,
- number of bytes packets received by the local host,
- TCP flags, especially SYN, RST and FIN flags
- duration of the connection.

Source/destination IP addresses and port number:

Changes in IP address and port number space are measured by observing the value of Shannon entropy related to that [8]. Entropy (a numerical measure of the uncertainty of an outcome) values are calculated based on different interval of time or for separate time periods. The length of the period can be a subject of more detailed discussion [8], however we assume that it is possible that different monitoring agents (MoA) use various periods length. This means that we will evaluate, collect and investigate the following network variables:

- S_IP (t_i) - entropy of source IP address in the period t_i,
- D_IP(t_i) - entropy of destination IP address in the period t_i,
- D_Port(t_i) - entropy of destination port number in the period t_i ,
- S_Port(t_i) - entropy of source port number in the period t_i .

Entropy value is evaluated from standard formula:

$$e = -\sum_{i=0}^{N} p_i \log p_i , \quad p_i = \frac{n_i}{\sum_{i=0}^{N} n_i}, 0 \le i \le N,$$ where:

N - cardinal number of IP address/port number set,
n_I - number of packets with a particular source/destination IP address/port number observed in the period t_I,
$\sum_{i=0}^{N} n_i$ - total number of packets observed in the period t_I.

As for some t_i , the value of $\sum_{i=0}^{N} n_i$ can be equal to zero (no traffic observed in t_i period), we assume that in these periods entropy value is also zero.

Any untypical changes of variables values related to IP address or port number entropy can be treated as a sign of anomalous behavior of the monitored system.

Especially we can assign some threshold value which will indicate the state of anomalous entropy level.

E.g. AS_IP will be a constant describing the value of acceptable S_IP level.

Number of bytes and packets:

Changes of entropy values are strictly related to changes of communication patterns. By using this measure, some sort of anomalies caused by intrusive actions like DoS or system scan can be detected.

However, other types of intrusions do not have to disturb communication patterns. For example so called topological worms using internally generated target lists tries to infect only well known by the infected host remote targets. Well known, means that instead of performing random scan to find vulnerable hosts, the worm tries to discover the local communication topology and infect only hosts which sent or received data to or form infected host [9].

The values describing number of bytes and packets exchanged by a host will be obtained as a result of observation of incoming and outgoing traffic in each of constant size period while it is observed by MoA.
TRAFFIC_B_R(t_i) - bytes received by a host in period t_i
TRAFFIC_B_S(t_i) - bytes sent by a host in period t_i
TRAFFIC_P_R(t_i) - packets received by a host in period t_i
TRAFFIC_P_S(t_i) - packets sent by a host in period t_i.
Also a traffic threshold value can be assigned and described.

TCP flags:

The TCP flags are important source of information about host's connections state. Typical TCP connection has three phases: connection establishment, data transfer, connection termination.

Each phase uses packets with some standard sequences of TCP flags; an especially TCP flag brings information about current connection state. However, this information may be incorrect while an intruder can manipulate the packet's content to reach some particular aim.

In our approach we measure a difference between number of sent SYN packets and received RST and FIN packets.

TCP-FLAG = $p_{t_i}^{syn} - p_{t_i}^{rst} - p_{t_i}^{fin}$ Where:

TCP_FLAG - parameter indicating temporal start/end connection ratio,

$p_{t_i}^{syn}$ - number of sent TCP packets with SYN flag set,

$p_{t_i}^{rst}$ - number of received TCP packets with RST flag set,

$p_{t_i}^{fin}$ - number of received TCP packets with FIN flag set. In normal conditions, in long time observation we should get the mean value of TCP_FLAG near zero. Intrusive actions like system scanning, DoS attacks may cause the temporal distortion of the mean value of TCP_FLAG.

Duration of the Connection:

Duration of a connection is one of attribute in attack detection process [7]. During various types of attacks, this value will be affected and so an anomaly may be detected. We evaluate simple mean value of connections duration that have been observed in period t_i.

c_{t_i} - mean value of duration of connections that have been observed in a period t_i.

D. Attack Pattern Ontology

Gruber [11] defines ontology as an explicit specification of conceptualization. The term, which is borrowed from philosophy, is used to provide a formal specification of the concepts and relationships that can exist between entities within a domain.

This paper proposes the core ontology which is containing basic concepts for defining attack patterns, in order to help in defining attacks and to simplify network variable based computations. The ontology contains basic concepts like *Attack* which is characterized by certain *Attack Pattern*, which in turn is defined by certain set of observations of the network variables.

As mentioned above, the observations are given in form of MoA's communicates about probability of anomalous variable value. Our ontology contains also specific operators which allow to define a *sequence* (SEQ) of communicates, their *concurrency* (AND) or *alternative* (OR). It is also possible to consider *paths* (i.e. sequences

of nodes) in network graph (PATH) and origin-destination pairs of network nodes (PAIR).

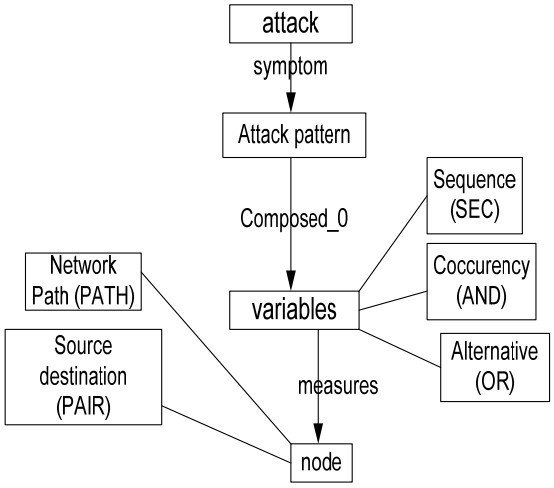


Figure 3: Core attack ontology

The following generic form of communicate about network variables are assumed:

MoA_1 (N_1, V_1) = x, where $x \in [0,1]$ Which should be

read: "*Monitoring agent MoA_1 states, that the value of network variable V_1 measured in node N_1 is normal (i.e. characteristic for the absence of attack) with probability x*". It is also assumed, that for any V_1 exists some threshold value A_i, such that any value MoA_1 (N_1, V_1) < Ai means that we experience an abnormal (suggesting that there's an attack) value of V1.

E. Attack Pattern Definition.

Let us look at a so called "*Trusted host*" Attack which can take place according to the following schemes:

In preparation of attack, the intruder prepares the attack by compromising several vulnerable hosts which create a network of so called "*zombie*" hosts.

Meanwhile, the intruder gathers information about the victim. This information will allow the attack to guess TCP sequence number of the victim.

At a specific time the intruder launch the attack and activates all "*zombie*" hosts to send spoofed SYN packets with the source address set to the victim's IP address to an agent "*trusted host*"**.**

The agent "*trusted host*" responds to this SYN packet by sending a SYN|ACK or a RST packet to the source address, which is actually the victim's IP address.

The victim replies with RST packets to *trusted host*'s SYN|ACK packets and with no packet to *trusted host*'s RST packets. The "*zombies*" send a continual storm of theses packets, thus causing the victim host to be flooded by innocent agent host ("trusted *host*").

Let us define this attack using ontology:

```
DEF_ATTACK (trusted host _Attack)
N1, N2, N3: Node;
//where N1-zombie, N2- victim, N3-
//Trusted host
EXISTS PATH(N1,N2) SUCH THAT:
(
SEQ (
// the sequence of the attack
( //
//Attack symptoms
MoA(N1, TCP_FLAG)<AFLAG_SYN
AND
MoA(N1, D_IP)<AD_IP
AND
MoA(N1, S_IP)<AS_IP
AND
MoA(N1, D_PORT)<AD_PORT
),
FOR ANY N3 in PATH(N1,N2)
// symptoms at "trusted host" nodes
(
MoA(N3, TCP_FLAG)<ATCP_FLAG
AND
MoA(N3, D_PORT)<AD_PORT
),
(
// symptoms at "victim"
MoA(N2, TCP_FLAG )<ATCP_FLAG
AND at
MoA(N2, TRAFFIC_B_R)<ATRAFFI_B_R
)
)
```

In the above definition we use earlier defined the following variables:
D_IP, S_IP
D_PORT
TRAFFIC_B_R
TCP_FLAG

F. Reasoning About Attacks

This IDS realizes and alarms about intrusion according to MoA information and MA decisions. The accurateness of IDS decision depends on MoAs evaluation of observed values and MA ability to correctly recognize attack patterns using data delivered by MoA and attack pattern ontology. Reasoning about attack that is performed by IDS can be described by the following procedure.

1. Each MoA observes and evaluates a set of variables described above. During this step MoA updates variables values so they represent the current system state. Variable list obtained by MoA_1 may be similar to the following example:
- TCP_FLAG= 143
- D_IP= 3,21
- S_IP= 2,81
- D_PORT= 1,98
- TRAFFIC_B_R= 3962

2. During next step MoA estimates the abnormality level of collected values. After this step MoA will be able to

present its opinions about nodes states in a form of attack probabilities presented in section 5.

- MoA_1 (N_1, TCP_FLAG)= 0,143
- MoA_1(N_1 ,D_IP)= 0,21
- MoA_1(N_1,S_IP)= 0,81
- MoA_1(N_1,D_PORT)= 0,28

Where N_1 – the node observed by MoA_1.

3. Third step performed by MoA is a comparison of current probability attack value with corresponding threshold value. For example:

MoA_1(N_1,TCP_FLAG)<0,05

- MoA_1(N_1,D_IP)<0,01
- MoA_1(N_1,S_IP)<0,5
- MoA_1(N_1,D_PORT)<0,3

As the result, MoA gets some binary vector:

TRAFFIC B R	FLAG SYN	D-IP	S-IP	D-PORT
0	0	0	0	1

Where '0' value in a vector means 'normal state' and'1' stands for "anomaly".

4. Next step is performed by a MA. The MA collects and processes binary vectors obtained from MoAs. The MA compares vectors to the known attack patterns. For example the MA possess the following list of MoA binary vectors:

MoA (N_1, XXX)

TRAFFIC B R	FLAG SYN	D-IP	S-IP	D-PORT
1	1	0	0	0

MoA (N_2, XXX)

TRAFFIC B R	FLAG SYN	D-IP	S-IP	D-PORT
0	1	0	0	1

MoA (N_3, XXX)

TRAFFIC B R	FLAG SYN	D-IP	S-IP	D-PORT
0	1	1	1	1

MoA (N_4, XXX)

TRAFFIC B R	FLAG SYN	D-IP	S-IP	D-PORT
0	1	0	0	1

MoA (N_5, XXX)

TRAFFIC B R	FLAG SYN	D-IP	S-IP	D-PORT
0	0	0	0	1

Comparing the binary vectors to attack pattern defined in section 5 MA recognizes the presence of trusted host attack where node N_3 plays a role of zombie node, node N_4 and N_2 play roles of trusted hosts and node N_1 is a victim.

IV. CONCLUSION AND FUTURE WORK

In this paper an Agent based distributed intrusion detection system architecture has been presented.

The system is achieved by allowing each agent to be autonomous, responsible and cooperative according to work assigned.

A new method for attack detection based on attack pattern ontology has also been presented. The ontology helps to gather information from different source (MoA_s) and take decision about network security status.

For future work, the system presented is still being extended and implemented. Apart of the case of attack presented, they may be other several attack or abnormal situation in network system. So, during further steps of our work, we will reason about other scenario related to security events and enhance the addition proposal which correspond to the new scenarios.

REFERENCES

[1] J.P Anderson, Computer Security Threat Monitoring and Surveillance, tech, report, James P.Anderson Co, Fort Washington Pa., 1980.

[2] Axelsson, S., 2000. Intrusion Detection Systems: A Taxonomy and Survey. Tech. Rep. 99-15, Department of Computer Engineering, Chalmers of Technology , Göteborg, Sweden.

[3] White, M.G.B., Fisch, E. A., Pooch, U.W.: Cooperating security managers: A peer-based intrusion detection system. IEEE Network **10** (1996) .

[4]Spafford, E.H., Zamboni, D.: Intrusion detection using autonomous agents. Computer Networks **34** (2000) .

[5] C.A.Bolt, III David L. Fisher, Ph.D., Remote Agent Technology: An Approach to Monitoring and Testing Distributed Simulation Systems, Northrop Grumman Information Technology Defense Enterprise Solutions 12000 Research Parkway, Suite 132 Orlando, FL 32826-3211 boltco@northropgrumman.com .

[6] F. Bellifemine, G. Caire, D. Greenwood, Developing multi-agent systems with JADE .

[7] A. Beach, M. Modaff, Y.Chen, Network Traffic Anomaly Detection and Characterization. cs.northwestern.edu/~ajb200/anomaly%20detecion%20paper%201.0.pdf.

[8] C.E. Shannon, W. Weaver, The mathematical theory of communication, University of Illinois Press, Urbana, 1949.

[9] N.Weaver, V. Paxson, S.Staniford and R. Cunningham, A taxonomy of computer worms. ACM Workshop on Rapid Malcode - WORM '03, ACM Press, New York, NY, pp. 11-18, 2003.

[10] I. Kotenko, et al., Multi-Agent Modeling and Simulation of Distributed Denial-of-Service Attacks on Computer Networks. Proceedings of Third International Conference Navy and Shipbuilding Nowaday. St. Petersburg, pp. 38 47, 2003.

[11] T. F. Gruber. A Translation Approach to Portable Ontologies. Knowledge Acquisition,5(2):199–220, 1993.

[12] Gorodetski, V. I., Popyack, L. J., Kotenko, I.V., Skormin, V.A.: Ontology-based multi-agent model of information security system. In: 7th RSFDGrC. Number 1711 in Lecture Notes in Artificial Intelligence. Springer (1999) .

[13] Undercoffer, J., Pinkston, J.: Modeling computer attacks: A target-centric ontology for intrusion detection. In: CADIP Research

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 139-142

Formal Description and Analysis of Malware Detection Algorithm $\mathcal{A}_{MOM}$

Ying Zeng, Fenlin Liu, Xiangyang Luo, and Chunfang Yang
Information Science and Technology Institute, Zhengzhou, China
Email: zengying510@yahoo.com.cn

***Abstract*—Code obfuscation can alter the syntactic properties of malware byte sequences without significantly affecting their execution behaviors. Thus it can easily foil signature-based detection. In this paper, the ability of handling obfuscation transformations of the semantics-based malware detection algorithm $\mathcal{A}_{MOM}$ proposed by Gao et al. is discussed using abstract interpretation theory from a semantic point of view. First, a formal description of the algorithm $\mathcal{A}_{MOM}$ is proposed. Then an equivalent trace-based detector is developed. Finally, the oracle soundness and oracle completeness of the trace-based detector for a restricted class of obfuscation transformations which preserve the variation relation are shown.**

***Index Terms*—malware detection, code obfuscation, trace semantics, abstract interpretation**

I. Introduction

As the complexity of modern computing systems is growing, various bugs are unavoidable in software systems. This increases the possibility of the malware attack that usually exploits such vulnerabilities in order to damage the systems. Thus, the malware attack has become a serious threat in computer security, and therefore it is crucial to detect the presence of malicious code in software systems.

Nowadays, one of the most popular approaches to malware detection is signature-based detection [1]. In order to foil this detection, malware writers often use code obfuscation such as instructions reordering, semantics NOP insertion, and substitution of equivalent instructions to automatically generate metamorphic malware [2]. In fact, the majority of malware that appears today is a simple repacked version of old malware [3].

Different obfuscated versions of the same malware have to share (at least) the malicious intent, namely the maliciousness of their semantics, even if they might express it through different syntactic forms. Therefore, addressing the malware detection problem from a semantic point of view can lead a more robust detection system [4]. For example, Christodorescu et al. [5] put forward a semantics-aware malware detector $\mathcal{A}_{MD}$ that is able to handle some commonly used obfuscations, such as semantics NOP insertion, instructions reordering and so on. While Gao et al. [6] introduce another semantics-based malware detection algorithm $\mathcal{A}_{MOM}$ which is able to handle not only the obfuscations that $\mathcal{A}_{MD}$ can handle, but also some other obfuscations like the flattening obfuscation proposed by Wang et al. [7]. And this detection scheme can largely reduce the updating of virus definition databases. However, the authors did not give the specific obfuscations that $\mathcal{A}_{MOM}$ could handle, and discussed the ability of $\mathcal{A}_{MOM}$ handling obfuscation transformations using only the experiment results.

In this paper, a formal description of the semantics-based malware detection algorithm $\mathcal{A}_{MOM}$ proposed by Gao et al. is given from a semantic point of view. Then an equivalent trace-based detector D_{Tr} is constructed using abstract interpretation theory. Finally, the oracle soundness and oracle completeness of D_{Tr} have been shown for a restricted class of obfuscation transformations which preserve the variation relation.

II. Abstract Interpretation and Programming Language

A. Abstract Interpretation

Abstract interpretation [8] was originally developed by P. Cousot and R. Cousot as a general theory for designing and approximating the fixpoint semantics of programs. The basic idea is to approximate semantics obtained from computation on the concrete domain by substituting the concrete domains of computation and concrete semantic operations with abstract domains and corresponding abstract semantic operations. The concrete semantics of a program is computed on the concrete domain $\langle C, \leq_C \rangle$ which is a complete lattice, modeling the values computed by the program. The partial ordering $\leq_C$ models relative precision: $c_1 \leq_C c_2$ means that c_1 is more precise than c_2. Approximation is encoded by an abstract domain $\langle A, \leq_A \rangle$ which is also a complete lattice representing some approximation properties on concrete objects. The abstract semantics is computed on an abstract domain. Usually abstract domains are specified by Galois connections.

B. Programming Language

The language we consider is the simple imperative language introduced in Ref. [9]. The syntax of the language is given in Table I. The auxiliary functions in Table II are useful in defining the semantics of the considered programming language, which is described in Table III.

An environment $\rho \in \mathfrak{E}$ maps variables $X \in dom(\rho)$ to their values $\rho(X)$, so $\mathfrak{E} \triangleq \cup_{\mathcal{X} \subseteq \mathbb{X}} \mathfrak{E}[\![\mathcal{X}]\!]$, where $\mathfrak{E}[\![\mathcal{X}]\!] \triangleq \mathcal{X} \to \mathfrak{D}_\perp$ is the subset of environments ρ with domain $dom(\rho) \triangleq \mathcal{X}$.

AP-PROC-CS-09CN005

$\mathfrak{E}[\![P]\!]$ is the set of environments of a program P whose domain is the set of program variables: $\mathfrak{E}[\![P]\!] \triangleq \mathfrak{E}[\![var[\![P]\!]]\!]$. $\rho|_{\mathcal{X}}$, where $\mathcal{X} \subseteq \mathbb{X}$, is the restriction of environment ρ to the domain $dom(\rho) \cap \mathcal{X}$. Let $\rho[X := n]$ be the environment ρ where value n is assigned to variable X.

TABLE I.
THE SYNTAX OF THE SIMPLE IMPERATIVE LANGUAGE

Syntactic Categories:	Syntax:
$n \in \mathbb{Z}$ (integers)	$E ::= n \mid X \mid E_1 - E_2$
$X \in \mathbb{X}$ (variable names)	$B ::= true \mid false \mid$
$L \in \mathbb{L}$ (labels)	$E_1 < E_2 \mid \neg B_1 \mid B_1 \vee B_2$
$E \in \mathbb{E}$ (integer expressions)	$A ::= X := E \mid X := ? \mid skip \mid B$
$B \in \mathbb{B}$ (Boolean expressions)	$C ::= L_1 : A \to L_2; \mid$
$A \in \mathbb{A}$ (actions)	$L_1 : B \to \{L_T, L_F\};$
$C \in \mathbb{C}$ (commands)	$\mathbb{P} ::= \wp(\mathbb{C})$
$P \in \mathbb{P}$ (programs)	

Let $\mathfrak{F}[\![P]\!]$ denote the set of final states of program P, the set of finite maximal execution traces $\mathbf{S}^n[\![P]\!]$ can be defined as: $\mathbf{S}^n[\![P]\!] \triangleq \{\sigma \in \Sigma^n \mid n>0 \wedge \forall i \in [0,n-1] : \sigma_i \in \mathbf{C}(\sigma_{i-1}) \wedge \sigma_{n-1} \in \mathfrak{F}[\![P]\!]\}$, where Σ^n is the set of finite state sequences of length n. The maximal finite trace semantics $\mathbf{S}^+[\![P]\!]$ of program P is given as $\mathbf{S}^+[\![P]\!] \triangleq \cup_{n>0} \mathbf{S}^n[\![P]\!]$.

TABLE II.
AUXILIARY FUNCTIONS

Labels:	Variables:
$lab[\![L_1 : A \to L_2;]\!] \triangleq L_1$	$var[\![L_1 : A \to L_2;]\!] \triangleq var[\![A]\!]$
$lab[\![L_1 : B \to \{L_T, L_F\};]\!] \triangleq L_1$	$var[\![L_1 : B \to \{L_T, L_F\};]\!] \triangleq var[\![B]\!]$
$lab[\![P]\!] \triangleq \{lab[\![C]\!] \mid C \in P\}$	$var[\![P]\!] \triangleq \cup_{C \in P} var[\![C]\!]$
Action of a command:	Successors of a command:
$act[\![L_1 : A \to L_2;]\!] \triangleq A$	$suc[\![L_1 : A \to L_2;]\!] \triangleq L_2$
$act[\![L_1 : B \to \{L_T, L_F\};]\!] \triangleq B$	$suc[\![L_1 : B \to \{L_T, L_F\};]\!] \triangleq \{L_T, L_F\}$

A control flow graph $G=(V, E)$ is a graph with the vertex set V representing program commands, and edge set E representing control-flow transitions from one command to its successor. The control flow graph (CFG) can be easily constructed as follows:

- For each command $C \in \mathbb{C}$, create a CFG node $v_{lab[\![C]\!]}$ annotated with that command. Let $C[\![v]\!]$ denote the command at CFG node v.
- For each command $C \in \mathbb{C}$, $L_1 = lab[\![C]\!]$, for each label $L_2 \in suc[\![C]\!]$, create a CFG edge (v_{L_1}, v_{L_2}).

For a given CFG $G=(V, E)$, $entry(G)$ denotes the set of the entry vertexes. $Path(G)$ denotes the set of all paths in G. $Path_{mn}(G) = \{\theta \in Path(G) \mid \theta = v_m \to \cdots \to v_n\}$ denotes the set of all paths from node v_m to node v_n in G. Consider a path θ in the CFG from node v_1 to node v_k, $\theta = v_1 \to v_2 \to \cdots \to v_k$. There is a corresponding sequence of commands in program P, written $P|\theta = \{C_1, \ldots, C_k\}$, where $C_i = C[\![v_i]\!]$. Then we can express the set of possible states after executing the sequence of commands $P|\theta$ as $\mathbf{C}^k[\![P|\theta]\!](\rho, C_1)$.

TABLE III.
THE SEMANTICS OF THE SIMPLE IMPERATIVE LANGUAGE

Value Domains	Boolean Expressions $\mathbf{B} : \mathbb{B} \times \mathfrak{E} \to \mathfrak{B}_\perp$
$\mathfrak{B}_\perp \triangleq \{true, false, \perp\}$ (truth values) $n \in \mathbb{Z}$ (integers) $\mathfrak{D}_\perp$ (variable values) $\rho \in \mathfrak{E} \triangleq \mathbb{X} \to \mathfrak{D}_\perp$ (environments) $\Sigma \triangleq \mathfrak{E} \times \mathbb{C}$ (program states)	$\mathbf{B}[\![true]\!]_\rho \triangleq true$ $\mathbf{B}[\![false]\!]_\rho \triangleq false$ $\mathbf{B}[\![E_1 < E_2]\!]_\rho \triangleq \mathbf{E}[\![E_1]\!]_\rho < \mathbf{E}[\![E_2]\!]_\rho$ $\mathbf{B}[\![\neg B]\!]_\rho \triangleq \neg \mathbf{B}[\![B]\!]_\rho$ $\mathbf{B}[\![B_1 < B_2]\!]_\rho \triangleq \mathbf{B}[\![B_1]\!]_\rho \vee \mathbf{B}[\![B_2]\!]_\rho$
Arithmetic Expressions $\mathbf{E} : \mathbb{E} \times \mathfrak{E} \to \mathfrak{D}_\perp$	Actions $\mathbf{A} : \mathbb{A} \times \mathfrak{E} \to \wp(\mathfrak{E})$
$\mathbf{E}[\![n]\!]_\rho \triangleq n$ $\mathbf{E}[\![X]\!]_\rho \triangleq \rho(X)$ $\mathbf{E}[\![E_1 - E_2]\!]_\rho \triangleq \mathbf{E}[\![E_1]\!]_\rho - \mathbf{E}[\![E_2]\!]_\rho$	$\mathbf{A}[\![skip]\!]_\rho \triangleq \{\rho\}$ $\mathbf{A}[\![X := E]\!]_\rho \triangleq \{\rho \mid X := \mathbf{E}[\![E]\!]_\rho\}$ $\mathbf{A}[\![X := ?]\!]_\rho \triangleq \{\rho' \mid \exists z \in \mathbb{Z} : \rho' = \rho[X := z]\}$ $\mathbf{A}[\![B]\!]_\rho \triangleq \{\rho' \mid \mathbf{B}[\![B]\!]_{\rho'} = true \wedge \rho' = \rho\}$
Commands $\mathbf{C} : \Sigma \to \wp(\Sigma)$	
$\mathbf{C}(\langle \rho, L_1 : A \to L_2; \rangle) \triangleq \{\langle \rho', C' \rangle \mid \rho' \in \mathbf{A}[\![A]\!]_\rho \wedge lab[\![C']\!] = L_2\}$ $\mathbf{C}(\langle \rho, L_1 : B \to \{L_T, L_F\} \rangle) \triangleq \left\{ \langle \rho, C' \rangle \mid lab[\![C']\!] = \begin{cases} L_T & \text{if } \mathbf{B}[\![B]\!]_\rho = true \\ L_F & \text{if } \mathbf{B}[\![B]\!]_\rho = false \end{cases} \right\}$	

III. FORMAL DESCRIPTION OF MALWARE DETECTION ALGORITHM $\mathcal{A}_{MOM}$

The semantics-based malware detection algorithm $\mathcal{A}_{MOM}$ proposed by Gao et al. [6] compares the semantics of a program with the semantics of the malware to identify the malicious behavior in the program and detect whether the program is a variation of the malware with respect to a class of obfuscation transformations of which the specifications are given, i.e. $\mathcal{A}_{MOM}(P, M) = 1$.

In order to reason about the ability of the algorithm $\mathcal{A}_{MOM}$ handling obfuscation transformations, we give a formal description of the algorithm from a semantic point of view as follows. The algorithm proceeds in four steps:

1. Collect program invariants of program P and malware M by symbolic executions. Let $G=(V, E)$ be the CFG of a program P, $Path_{mn}(G) = \{\theta_1, \theta_2, \ldots, \theta_{num}\}$, where $num = |Path_{mn}(G)|$ is the size of set $Path_{mn}(G)$. The invariant at program point v_n can be expressed formally as $\varphi(v_n) = \bigvee_{i=1}^{num} \left(\left(\bigwedge_{X \in var[\![P]\!]} \left(X = \mathbf{E}[\![X]\!]_{\rho_i} \right) \right) \wedge \left(\bigwedge_{b \in B_i} b \right) \right)$, where $\rho_i = env[\![s_i]\!]$ is the environment in the state s_i after executing the sequence of commands $P|\theta_i$ from the initial state $\langle \rho_0, C[\![v_m]\!] \rangle$, $\theta_i \in Path_{mn}(G)$, B_i is the set of predicates needed to be satisfied while executing the sequence of commands $P|\theta_i$. This step makes use of two oracles:

OR_{CFG} that returns the control-flow graph $G^P=(V^P,\ E^P)$ of a program P and OR_{PI} that returns the invariants at all program points $\Psi^P=\{\varphi^P(v)\mid v\in V^P\}$ of a program P.

2. Identify a control-flow map and a data-flow map between malware M and program P according the specification of the obfuscation algorithm $\mathcal{O}$. There is a map matching a malware node v^M to a program node v^P, denoted by $\mu: V^M \to V^P$. And this map μ induces a map $\upsilon: var[\![M]\!]\times V^M \to var[\![P]\!]$ from variables at a malware node to variables at the corresponding program node.

3. Build an equivalent relation between the variables in the sets of nodes of the two CFGs. According to the specification of the obfuscation algorithm $\mathcal{O}$, the variable values in malware M should be equal to the corresponding variable values in program P. This equivalent relation, denoted by Q, can be expressed as $\forall v_k^M \in dom(\mu),\ X_k^M \in var[\![C[\![v_k^M]\!]]\!],\ \rho\in\mathfrak{E},\ s^M\in \mathbf{C}^k[\![M\mid \lambda^M]\!](\rho, C[\![v_0^M]\!])$: $\mathbf{E}[\![X_k^M]\!]_{env[\![s^M]\!]} = \mathbf{E}[\![\upsilon(X_k^M, v_k^M)]\!]_{env[\![s^P]\!]}$, where $v_0^M = entry(G^M)$, $s^P=\mathbf{C}^l[\![P\mid\lambda^P]\!](\rho, C[\![v_0^P]\!])$, $v_0^P = entry(G^P)$, $v_l^P=\mu(v_k^M)$, $\lambda^P=\mu_{path}(\lambda^M)$.

4. Check whether the equivalent relation built in step 3 holds. Construct a verification condition, formally described as $\bigwedge_{v_k^M\in dom(\mu)}\left(\varphi^M(v_k^M)\wedge\varphi^P(\mu(v_k^M))\right)\Rightarrow Q$. Pass this verification condition to a theorem prover. If the condition holds, then identify the program P as a variant of the malware M with respect to the obfuscation algorithm $\mathcal{O}$, i.e. $\mathcal{A}_{MOM}(P,M)=1$. This check is implemented in $\mathcal{A}_{MOM}$ as a query to oracle $OR_{Validation}$, which determines whether a verification condition holds.

IV. An Equivalent Trace-Based Detector D_{Tr}

In the following, we first give the definitions of three abstractions α_{MOM}, α_{env} and α_r.

The abstraction α_{MOM}, when applied to a trace $\sigma\in\mathbf{S}^+[\![P]\!]$, with $\sigma=(\rho_1',C_1')\cdots(\rho_n',C_n')$, to a set of variable maps $\{\pi_i\}$, and a set of location maps $\{\gamma_i\}$, returns an abstract trace $\alpha_{\mathrm{MOM}}(\sigma,\{\pi_{\mathrm{i}}\},\{\gamma_i\})=(\rho_1,\mathrm{C}_1)\cdots(\rho_{\mathrm{n}},\mathrm{C}_{\mathrm{n}})$, if $\forall i, 1\le i\le n$, $act[\![C_i]\!]=act[\![C_i']\!][X/\pi_i(X)]$, $lab[\![C_i]\!]=\gamma_i(lab[\![C_i']\!])$, $suc[\![C_i]\!]=\gamma_i(suc[\![C_i']\!])$, $\rho_{\mathrm{i}}=\rho_{\mathrm{i}}'\circ\pi_{\mathrm{i}}^{-1}$, where $A[X/\pi(X)]$ represents actions A where each variable name X is replaced by the new name $\pi(X)$. Otherwise, if the condition does not hold, then $\alpha_{\mathrm{MOM}}(\sigma,\{\pi_{\mathrm{i}}\},\{\gamma_{\mathrm{i}}\})=\varepsilon$. A map $\pi_i: var[\![P]\!]\to var[\![P']\!]$ renames program variables $var[\![P]\!]$ such that they match program variables $var[\![P']\!]$, $\gamma_i: lab[\![P]\!]\to lab[\![P']\!]$ reassigns program memory locations $lab[\![P]\!]$ to program memory locations $lab[\![P']\!]$.

Given a trace $\sigma=(\rho_1,C_1)\sigma'$, the abstraction α_{env} retains only the environments,

$$\alpha_{env}(\sigma)\triangleq\begin{cases}\varepsilon & \text{if } \sigma=\varepsilon\\ \rho_1\alpha_{env}(\sigma') & \text{if } \sigma=(\rho_1,\mathrm{C}_1)\sigma'\end{cases}. \quad (1)$$

Let $lab_r[\![P]\!]\subseteq lab[\![P]\!]$ be a restriction of a program P, the abstraction α_r propagates the restriction $lab_r[\![P]\!]$ on a given trace $\sigma=(\rho_1,C_1)\sigma'$ as

$$\alpha_r(\sigma, lab_r[\![P]\!])\triangleq\begin{cases}\varepsilon & \text{if } \sigma=\varepsilon\\ (\rho_1^{\,r},C_1)\alpha_r(\sigma') & \text{if } lab[\![C_1]\!]\in lab_r[\![P]\!]\\ \alpha_r(\sigma') & \text{otherwise}\end{cases}, \quad (2)$$

where $\rho_1^{\,r}\triangleq\rho_{1|var_r[\![P]\!]}$, $var_r[\![P]\!]\triangleq\cup\{var[\![C]\!]\mid lab[\![C]\!]\in lab_r[\![P]\!]\}$.

We can model the algorithm $\mathcal{A}_{MOM}$ using these three abstractions α_{MOM}, α_{env} and α_r. The abstraction α that characterizes the trace-based detector D_{Tr} is given by the composition of these three abstractions $\alpha_{env}\circ\alpha_{MOM}\circ\alpha_r$. We will show that D_{Tr} is equivalent to $\mathcal{A}_{MOM}$, when the oracles it uses are perfect.

Definition 1: Malware detector D_{Tr} is an α-semantic malware detector defined on the abstraction α, it classifies a program P as a variation of a malware M, i.e. $D_{Tr}(\mathbf{S}^+[\![P]\!],\mathbf{S}^+[\![M]\!])=1$, if

$$\begin{aligned}&\exists lab_r[\![P]\!]\in\wp(lab[\![P]\!]), lab_r[\![M]\!]\in\wp(lab[\![M]\!]),\\ &\{\pi_i: var[\![P]\!]\to var[\![M]\!]\}_{i\ge1},\{\gamma_i: lab[\![P]\!]\to lab[\![M]\!]\}_{i\ge1}:\\ &\alpha(\mathbf{S}^+[\![M]\!], lab_r[\![M]\!],\{\pi_i\},\{\gamma_i\})=\alpha(\mathbf{S}^+[\![P]\!], lab_r[\![P]\!],\{\pi_i\},\{\gamma_i\})\end{aligned}, \quad (3)$$

where $\alpha=\alpha_{env}\circ\alpha_{MOM}\circ\alpha_r$.

Proposition 1: The semantics-based malware detection algorithm $\mathcal{A}_{MOM}$ is equivalent to the $\alpha_{env}\circ\alpha_{MOM}\circ\alpha_r$-semantic malware detector D_{Tr}, i.e.

$$\forall P,M\in\mathbb{P}: \mathcal{A}_{MOM}(P,M)=1\Leftrightarrow D_{Tr}(\mathbf{S}^+[\![P]\!],\mathbf{S}^+[\![M]\!])=1. \quad (4)$$

One of the most important requirements of a robust malware detection algorithm is to handle obfuscation transformations. For a malware detector, this can be formalized in terms of soundness and completeness properties [4]. Intutitively, a malware detector is sound if it never erroneously claims that a program is infected (no false positives) and it is complete if it always detects program that are infected (no false negatives). When a program P is a variation of a malware M with respect to an obfuscation $\mathcal{O}$, it can be denoted by $P\approx\mathcal{O}(M)$. A malware detector D is sound (complete) for an obfuscation $\mathcal{O}\in\mathbb{O}$ if and only if

$$\forall M,P\in\mathbb{P}: D(P,M)=1\Rightarrow P\approx\mathcal{O}[\![M]\!]\,(P\approx\mathcal{O}[\![M]\!]\Rightarrow D(P,M)=1). \quad (5)$$

Most malware detectors are built on top of other static analysis techniques for problems that are hard or undecidable. So Ref. [4] introduced the notion of relative sound-

ness and completeness with respect to algorithms that a detector uses. A malware detector $D^{\mathcal{OR}}$ is oracle sound (complete) with respect to an obfuscation $\mathcal{O}$, if $D^{\mathcal{OR}}$ is sound (complete) for that obfuscation when all oracles in the set $\mathcal{OR}$ are perfect.

Following we define a class of obfuscations $\mathbb{O}_{MOM}$ which preserve the variation relation, namely, there is an variation relation between the original program M and obfuscated program $\mathcal{O}(M)$, formally expressed as

Definition 2: The obfuscation $\mathcal{O} \in \mathbb{O}_{MOM}$ preserves variation relation, if $\forall M \in \mathbb{P}$, such that

$$\begin{aligned} &\exists lab_R[\![M]\!] \in \wp(lab[\![M]\!]), lab_R[\![\mathcal{O}(M)]\!] \in \wp(lab[\![\mathcal{O}(M)]\!]), \\ &\{\xi_i : var[\![\mathcal{O}(M)]\!] \to var[\![M]\!]\}_{i\geq 1}, \{\vartheta_i : lab[\![\mathcal{O}(M)]\!] \to lab[\![M]\!]\}_{i\geq 1} : \\ &\alpha_{env}\left(\alpha_{MOM}\left(\alpha_r\left(\mathbf{S}^+[\![M]\!], lab_R[\![M]\!]\right), \{\xi_i\}, \{\vartheta_i\}\right)\right) \\ &= \alpha_{env}\left(\alpha_{MOM}\left(\alpha_r\left(\mathbf{S}^+[\![\mathcal{O}(M)]\!], lab_R[\![\mathcal{O}(M)]\!]\right), \{\xi_i\}, \{\vartheta_i\}\right)\right) \end{aligned} \quad (6)$$

Therefore, the following properties can be easily obtained, showing that the malware detector D_{Tr} which is equivalent to $\mathcal{A}_{MOM}$ is oracle sound and oracle complete with respect to this class of obfuscations.

Property 1: Malware detector D_{Tr} is oracle sound for the obfuscation $\mathcal{O} \in \mathbb{O}_{MOM}$, i.e.

$$D_{Tr}\left(\mathbf{S}^+[\![P]\!], \mathbf{S}^+[\![M]\!]\right) = 1 \Rightarrow \exists \mathcal{O} \in \mathbb{O}_{MOM} : P \approx \mathcal{O}(M). \quad (7)$$

Property 2: Malware detector D_{Tr} is oracle complete for the obfuscation $\mathcal{O} \in \mathbb{O}_{MOM}$, i.e.

$$\exists \mathcal{O} \in \mathbb{O}_{MOM} : P \approx \mathcal{O}(M) \Rightarrow D_{Tr}\left(\mathbf{S}^+[\![P]\!], \mathbf{S}^+[\![M]\!]\right) = 1. \quad (8)$$

V. Conclusions

The semantics-based malware detector $\mathcal{A}_{MOM}$ proposed by Gao et al. can detect whether a program is a variation of a malware with respect to some commonly used obfuscations, and largely reduce the updating of malware definition databases. In this paper, a formal description of the malware detection algorithm $\mathcal{A}_{MOM}$ proposed by Gao et al. is given from a semantic point of view and an equivalent trace-based detector D_{Ir} is constructed by abstract interpretation. At last, the oracle soundness and oracle completeness of the detector D_{Ir} for a restricted class of obfuscation transformations which preserve the variation relation have shown. Our work provides a formal basis for addressing the ability of the malware detection algorithm $\mathcal{A}_{MOM}$. The properties of $\mathcal{A}_{MOM}$ such as soundness and completeness can be proved using the result of this paper in the framework proposed by Preda et al. [4] which can be used to reason about and evaluate the resilience of malware detectors to various kinds of obfuscation transformations. Also our work can be a reference for designing effective malware detection algorithms.

Acknowledgment

The authors wish to thank X. Y. Luo, C. F. Yang, and L. Bin for comments and suggestions. This work was supported in part by grants from the National Natural Science Foundation of China under Grant Nos. 60970141, 60902102.

References

[1] P. Szor, *The Art of Computer Virus Research and Defense*. Boston, MA: Addison-Wesley Professional, 2005.

[2] C. Nachenberg, "Computer virus-antivirus coevolution," *Comm. ACM*, vol. 40(1), pp. 46–51, 1997.

[3] R. Perdisci, A. Lanzi, and W. Lee, "Classification of packed executables for accurate computer virus detection," *Pattern Recognition Letters*, vol. 29(14), pp. 1941–1946, 2008.

[4] M. D. Preda, M. Christodorescu, S. Jha, and S. Debray, "A semantics-based approach to malware detection," *ACM Trans. Program. Lang. Syst*, vol. 30(5), pp. 1–54, 2008.

[5] M. Christodorescu, S. Jha, S. A. Seshia, D. Song, and R. E. Bryant. "Semantics-aware malware detection," In *Proceedings of the 2005 IEEE Symposium on Security and Privacy (S&P'05)*, IEEE Computer Society, pp. 32–46, 2005.

[6] Y. Gao, Y. Y. Chen, and B. J. Hua, "A semantics-based malware detector for obfuscated malware," *Journal of Chinese Computer Systems*, vol. 28(1), pp. 1–8, 2007.

[7] C. X. Wang, "A security architecture for survivability mechanisms," *PhD Dissertation*, University of Virginia, Department of Computer Science, 2000.

[8] P. Cousot, and R. Cousot, "Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fixpoints," In *Proceedings of the 4th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL'77)*, ACM Press, pp. 238–252, 1977.

[9] P. Cousot, and R. Cousot, "Systematic design of program transformation frameworks by abstract interpretation," In *Proceedings of the 29th ACM SIGPLAN-SIGACT symposium on Principles of Programming Languages (POPL'02)*, ACM Press, pp. 178–190, 2002.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 143-146

On-line Modeling Method of Rolling Mill Based Least-squares Regression Analysis

Zhou Wan , Xiaodong Wang , Xin Xiong , and Jiande Wu
Faculty of Information Engineering and Automation
Kunming University of Science and Technology,Kunming, China
Email: ynkgwz@yahoo.com.cn

***Abstract*—Iron and steel industry is the base of the country economy. It is the important support of the social developing, and is the important basic industry of the modern economic. This paper studies the rolling mechanical model of rolling process control, fluid stress model, rolling torque model and the temperature change models. And put forward an on-line modeling method of rolling mill based Least-squares regression analysis, this method improve the accuracy of mathematical model of rolling. In a word, steel rolling is complicated system engineering. The system optimization with the system engineering to analyze the whole produces process is an inevitable trend.**

***Index Terms*—rolling mill; on-line model; least-squares regression analysis**

I. INTRODUCTION

Iron and steel industry is raw material industry, the national economy in the basic industries. Among them, steel rolling technology is the key working procedure in the Iron and steel production process. This is one area which is the fastest growing in the steel industry. And a variety of new technologies are the most widely used in this area. The primary problem of mill control is mathematical model. In the rolling process, the flow stress model, rolling force model, torque model and temperature change models are the core models of impacting rolling force changing. To enable smooth passage of the steel rolling line, to get a finished product of a certain dimension precision, good plate type, having a certain mechanical property, and to ensure that the entire rolling line equipment safety in the production process is necessary to establish a complete set of computers that can be used by the computer.

In actual production, only the mechanism model or empirical model is not enough, these models must be converted on-line model which can be used by computer through appropriate methods, which can be used by computer. This paper is adoption of the method of least-squares multivariate regression analysis to establish an on-line model of rolling mill.

The paper is supported by the foundation of Yunnan Education Department (09Y0094) and foundation of Kunming University of Science and Technology (KKZ1200903011).

II. ROLLING OF ROLLING MILL ANALYSIS

Rolling mill production process includes the continuous casting slab transmission, embryonic plate heating by walking beam furnace, dephosphorizing by high-pressure water, Rolling of rolling mill (including the width of vertical rolling controlling), laminar flow refrigeration, coiling on the ground, bundling, weighing, marking and hot rolled coil cooling. Its products available for cold rolling mill, or hot-rolled product volumes, or processed into hot-rolled plates through cross-cutting units, the joint shear unit.

In the rolling process, according to the different range of controlled object, the rolling process mathematical model can be a single formula can also be a combination of a group of formula [1], which involves a number of non-linear model.

The models have a greater impact on rolling force changing, including flow stress model, rolling force model, rolling torque model. The following analysis is on these major models.

A. The mechanical model of rolling force

In the rolling process, the rolling pressure described as: there are two metal-to-roll acting force: one is the resultant force of tangent to the contact surface friction stress- friction; the other is the resultant force of the roller and the rolling contact surfaces perpendicular to the unit pressure- positive pressure. The projection sum of the friction and positive pressure projection in the perpendicular to the rolling direction - vertical force of parallel roller center to connect, is often called rolling pressure, sometimes called the rolling force. Usually refers to the actual measured total pressure which measured with the pressure measuring instrument. The magnitude of the rolling pressure depends on the value of unit pressure and its distribution along the contact arc characteristics.

Rolling pressure is the most important force and energy parameters of proper design and rational use of rolling mill, but also is the main process parameters and device parameters. Use it to allocate sub channel load, develop technology systems, adjustment of rolling mill and strengthening rolling, expanding range of products and equipment to fully tap the potential in a reasonable manner. At the same time through the rolling process can

determine the motor power, checking and calculating the strength and deformation of various parts of the parts of rolling mill. In the computer-controlled rolling process, the rolling force model has always occupied an important position, it plays an extremely important role in the rolling process computer control technology. The forecasting accuracy of rolling force model is not only having direct impact on setting accuracy, but also having direct impact on the thickness and shape and so on. Simultaneously, as the rolling process to obey the principle of equal flow of seconds [2], thickness control improper will result in flow imbalance, making the process characteristics deviate from the set state, sequentially, having a negative affect on the stability of the process and working conditions of the regulatory system.

B. Flow Stress Model

External force applied to an object, the object always generates internal forces inside the object. Its internal force generation is used to balance the external force. This internal force distributed on the unit area is called stress, namely:

$$\sigma = \frac{P}{F} \tag{1}$$

Where σ —stress
P—pressure
F—force size

Metal plastic deformation is, under the external force, metal plastic deformation is non-synchronous mobile process of a large number of metal atoms, from some stable equilibrium position to some other stable equilibrium position. This process must be under a certain stress field to overcome the elastic force to completed, and this elastic force can make metal atoms trying to return to the original stable equilibrium position. The mechanical index which measures capability of object to maintain its original shape and resist deformation, is defined as metal plastic deformation resistance. The metal plastic deformation resistance refers to, in the units of stress state, the force per unit which is needed by metal materials to produce plastic deformation. Its size not only related to materials and components, but also related to the physical conditions of plastic deformation (deformation temperature, deformation speed and deformation degree) [3].

As the deformation resistance model is an important physical parameter of rolling pressure calculation formula, is an important model of affecting the accuracy of rolling pressure. Thus, decades, many scholars dedicated to experimental research work of metal plastic deformation resistance, made some useful data. Studying up on the mathematical model of metal plastic deformation resistance mainly focused on improving and optimizing the mathematical model [4-9].

Theoretically, the relationship between deformationv resistance and its influence factors can be expressed as the following functional form:

$$k_t = e^{At+B} \tag{2}$$

Where $x\%$ for the chemical composition and organization Status; t for the deformation temperature; $\dot{\varepsilon}$ for the deformation rate; ε for the deformation degree; τ for the deformation history effects; k_τ —Deformation history for deformation resistance of the influence coefficient.

C. Rolling torque model

Once the rolling pressure is known, you can find the rolling torque. Rolling torque is the product of rolling pressure and arm of force [10-11].

When rolling is not subject to other external force, according to Rolling Theory, could calculate to get the rolling moment as follows:

$$T = 2Pa = 2Pl_c\varphi \tag{3}$$

In formula P — rolling pressure
a — arm of force
φ — arm coefficient

During the hot rolling, arm coefficient φ =0.39～0.48, of which:

Roughing mill group: φ =0.44～0.48

Finishing mill group: φ =0.39～0.44

Thus it can be seen, in the method of adopting rolling pressure to calculate the rolling arm of force, the main difficulty is how to correctly determine the arm coefficient φ. To determine the φ value at present there are still large errors. Therefore, we generally use the empirical method when calculating the rolling torque.

It is described as the product of rolling pressure, roll contact arc length and leverage factor. Adopting the following approach:

$$T = P \cdot \sqrt{R'\Delta h} \cdot L_a \tag{4}$$

In the formula, the La—scale factor, usually made reference to leverage factor, and describe geometric interdependency of torque.

After determining leverage factor of different sequences in Plate Rolling by Sims, functional dependency of leverage factor is given by formula (5), it uses the roller occlusal contact area arc L and the average thickness of h.

$$L_a = Ae^{\frac{B \cdot L}{h}} \tag{5}$$

Among them, two parameters A and B are self-adapting, they are acquired through the most suitable conditions for the measured torque by self-learning. This ensures that the friction coefficient is a constant, along this roll contact zone. Leverage factor expressed torque interdependency in roll occlusal area. This factor has been modeled as a function which described by length of roller contact area, exit thickness and two self-adapting parameter. For each steel type models,

The torque model and the calculation of leverage factor setting should be kept separate. Through long-term

self-learning, they are able to adjust these coefficients automatically based on the actual measured data.

III. ESTABLISHED METHOD OF ON-LINE MODEL METHOD BASED LEAST-SQUARES

Taking full account of the premise of the necessary conditions, when modeling, we use multiple linear regression analysis method:

According to the above method of theoretical physics, method of experimental physics or according to the field experience (from all the relevant parameters, selected a number of factors that play a major influence as independent variables) set up a static mathematical model (prediction equation) in general can be summed up as follows:

$$y = a_0 + a_1 x_1 + a_2 x_2 \ldots + a_m x_m + \Delta \tag{6}$$

Where y - dependent variable;

$x_1, x_2, \ldots, x_m$ —m independent variables(main factors);

$a_0, a_1, \ldots, a_m$ —m +1 factor of the model;

Δ —model error

The establishment of a static mathematical model basing on specific structure of $y, x_1, x_2, \ldots, x_m$, that identified on applying the all above methods, collecting a large number of measured data (after sieving)at the scene, applying statistical methods to derive $a_0, a_1, \ldots, a_m$,the quantitative value of each coefficient, according to a certain independent variables $x_1, x_2, \ldots, x_m$ to accurately predict the value of y .

Measured data collected at the scene, in addition to y and its corresponding a group of $x_1, x_2, \ldots, x_m$ data measured record must be complete, still should record the information to determine the normal production process of the relevant quantitative and qualitative information at the same time. This can be used as the basis for deciding whether to delete while disposing data (In the data processing will also remove some data with too much errors through the precision analysis).If the data is not enough after deleting, the actual measured data should be added.

It should be noted is that the actual mathematical model structure directly perceived through the senses is different from the formal one, and generally speaking, may not certain be linear. But in normal condition, they all can be attributed to the above formula form through some type of transformation. Therefore, introducing linear multi-equation for manipulation data has universal significance.

Collecting large amounts of data and after deleting, then get a group (n group) data:

$$y_i, x_{1i}, x_{2i}, \ldots, x_{mi} \qquad (i = 1 \sim n) \tag{7}$$

On this basis, the following equations are:

$$\begin{aligned} y_1 &= a_0 + a_1 x_{11} + a_2 x_{21} + \ldots + a_m x_{m1} + \Delta_1 \\ y_2 &= a_0 + a_1 x_{12} + a_2 x_{22} + \ldots + a_m x_{m2} + \Delta_2 \ \ldots\ldots\ldots\ldots\ldots \\ y_n &= a_0 + a_1 x_{1n} + a_2 x_{2n} + \ldots + a_m x_{mn} + \Delta_n \end{aligned} \tag{8}$$

Where. $n >> (m+1)$

Therefore, We got a group of "redundant equations"--the number of equations is far greater than the number of unknowns, furthermore, these equations are all with the error. We need to establish an optimal criterion, as using it to determine the "optimal" m +1 unknowns from the extra equations. Commonly we use the "minimum squares sum of error term" optimal criteria to deal with these equations. It is generally known as the least-squares regression analysis method (multiple regression).

Minimum squares sum of error term is:

$$J = \sum_{j=1}^{n} \Delta_j = \sum_{j=1}^{n} \left[y_j - (a_0 + a_1 x_{1j} + a_2 x_{2j} + \ldots + a_m x_{mj}) \right]^2 = \text{Min} \tag{9}$$

necessary condition of the function

From the necessary condition for extremal functions, we can obtain m +1 equations, namely,

$$\begin{aligned} &\frac{\partial J}{\partial a_0} = 0 \\ &\frac{\partial J}{\partial a_1} = 0 \\ &\ldots\ldots\ldots \\ &\frac{\partial J}{\partial a_m} = 0 \end{aligned} \tag{10}$$

Thus available

$$\frac{\partial J}{\partial a_0} = -2 \sum_{j=1}^{n} \left[y_j - (a_0 + a_1 x_{1j} + a_2 x_{2j} + \ldots + a_m x_{mj}) \right] = 0 \tag{11}$$

Can be solved

$$n a_0 = \sum_{j=1}^{n} y_i - a_1 \sum_{j=1}^{n} x_{1j} - a_2 \sum_{j=1}^{n} x_{2j} - \ldots - a_m \sum_{j=1}^{n} x_{mj} \tag{12}$$

So

$$a_0 = \bar{y} - a_1 \bar{x}_1 - a_2 \bar{x}_2 - \ldots - a_m \bar{x}_m \tag{13}$$

Where

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{n} y_j \quad , \bar{x}_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij} \qquad (i = 1 \sim m) \tag{14}$$

Then $(i = 1 \sim m)$ (15)

Obtained m equations, substituted into a_0 , m unknowns can be solved, which is $a_1, \ldots, a_m$, and then substituted $a_1, \ldots, a_m$ into the formula of a_0 can be derived results of a_0 . This is the optimal estimate value of model coefficients in the least-squares method[11].

IV. CONCLUSION

In the actual production, only using mechanism model or empirical model is not enough. These models will be converted into the on-line model applied to computer. The main models of rolling process control are: Rolling mechanical model, fluid stress model, rolling torque model and the temperature change model, and these major models are analyzed, in this paper. Put forward an on-line modeling method of rolling mill based Least-squares regression analysis, this method improve the accuracy of mathematical model of rolling

REFERENCES

[1] X. Wang, The integrated use of artificial neural networks and teaching model in the Hot Rolling Mill Rolling Force Prediction, iron and steel, 1999, (3) :37-40.

[2] RB. Sims, The calculations of roll force and torque in hot rolling, Proceedings of the Institution of Mechanical Engineers, 1954, (168): 191-200.

[3] F., Yamada, Sekiguchi, K., Tsugeno, M., Anbe, Y., Hot strip mill mathematical models and set-up calculation, IEEE Transactions on Industry Application, 1991, 27:131-139.

[4] AK, Ghosh, A physically-based constitutive model for metal deformation, Acta. Metal., 1980, 28: 1443-1465.

[5] Hatta, N. Kokado, JI, Kikuchi, S., Takuda, H., Modeling on flow stress of plain carbon steel at elevated temperatures, Steel Res. 1985, 56 (11): 575-582.

[6] A., Laasraoui, Jonas, JJ, Prediction of flow stresses at high temperatures and strain rates, Metall. Trans., A. 1991, 22A (7) :1545-1558.

[7] H., Takuda, Fujimoto, H., Hatta, N, Modeling on flow stress of Mg-Al-Zn alloys at elevated temperatures. Journal of Materials Processing Technology, 1998, 80-81: 513-516.

[8] I., Schindler, Hadasik, E., A new model describing the hot stress-strain curves of HSLA steel at high deformation. Journal of Materials Processing Technology, 2000, 106: 131-135.

[9] S., Serajzadeh, Taheri, AK, Prediction of flow stress at hot working condition. Mechanics Research Communications 2003, 30: 87-93.

[10] Y. Sun, The basis of mathematical model of hot strip, Beijing, Metallurgical Industry Press, 1979.

[11] X. Wang, Optimize Control of Model Adaptive for Two-Stand Tandem Reversing Steckel Mill, Ph. D. thesis, Kunming University of Science and Technology, Kunming, 2009

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 147-150

Based on Exponential Smoothing Model of the Mill Self-learning Optimization Control

Xin Xiong , Xiaodong Wang, Zhou Wan , and Jiande Wu
Faculty of Information Engineering and Automation
Kunming University of Science and Technology,Kunming, China
Email: kmxiongxin@gmail.com

***Abstract*—Steel rolling process control is the key working procedure of the steel yield process, and that the control of the rolling process is the important factor to affect the quality and cost of the product. This paper is focus on the mill control. Firstly, the model of the mill is founded by exponential smoothing. Then, the method of self-learning optimization can be used to control the model. The online model, model adaptive learning control method, and adaptive learning optimize method can be founded. The model adaptive of rolling control model, careless rolling, and extract rolling can be solved greatly. The control effect and and good performance are obtained.**

***Index Terms*—exponential smoothing; self-learning; optimization; mill**

I. INTRODUCTION

Iron and steel industry is the foundation of national economic and the modern economy. It is also an important pillar of social development of the industry . The first question is rolling mill control mathematical model which is use theory method to establish. both must assume some conditions, also ignore some conditions, thus model to forecast the error. Due to the limited of the test methods and data distribution, the model of the statistical method to establish is inevitable meeting produces error. In the actual hot-rolling production process, no mathematical model can achieve 100% throughout the calculation precision of the mathematical model. Besides itself, the error of measurement, the change of characteristics and the uncertainty of material of rolling mill systems also affects model accuracy. Good condition of equipment and technology is the prerequisite and the necessary conditions to improve the accuracy of model system. The mathematical model of online adaptive fixed (learning) is the effective method to improve mathematical model [1-4].

The purpose of learning is according to the change of state system. In order to ensure the precision of the model, it use real-time information to amend the parameters of the model. Through some of the adaptive algorithm and using the reliable data of production process, amend the mathematical model of the relevant parameters, or the mathematical model of the adaptive correction coefficient, real-time online, namely "adaptive correction". The current application model adaptive correction method has many kinds. The paper is based on fluid for stress model put forward self-learning optimization control method of based on exponential smoothing[5].

The paper is supported by the foundation of Yunnan Education Department (09Y0094) and foundation of Kunming University of Science and Technology (KKZ1200903011).

II. MODEL SELF-LEARNING METHOD ANALYSIS

A. Exponential Smoothing

The learning method use exponential smoothing to study. The weighted average index of time sequence is the core of exponential smoothing. Thus, due to the situation of the time sequence has stability and regularity, so it can be postponed reasonably and conveniently. Recently, in the past to some extent recently, so will the future of weight in recent information. Not only abandoned past data, but only to gradually diminishing influence, pay more attention to the latest figures[6-7].

In order to reduce the workload of the adaptive correction calculation, the mathematical models can be expressed under the:

$$y = a_1 x_1 + a_2 x_2 \ldots + a_m x_m + \beta \tag{1}$$

Where $x_1, x_2, \ldots, x_m$ have a direct impact on the model of factors. $a_0, a_1, \ldots, a_m$ Indicate that the role of these factors have influence to Y on the degree. Because it is a linear model, the impact of various factors can be superimposed. And $a_0, a_1, \ldots, a_m$ will change along with the characteristics of the system changes with the change over time. In order to simplify the problem, self-learning calibration in the model calculations, can be made by a number of data model parameters: $a_0, a_1, \ldots, a_m$ as the inherent characteristics of that system to take the volume of fixed value. Change in the system state is expressed with the style β .When the state of the system generated change, the model coefficients can be amended accordingly calculated to meet the changes in system properties. According to equation (1) the concept of near real-side data can be projected out of the value of the counter-part of the information to correct model.

B. The model of self-learning process

An arbitrary linear model function with the following formula is set by[7]:

$$Y = f(X_{1,} X_{2,} \cdots X_m) \tag{2}$$

Where Y—the output variables of mathematical model;
X—input variables of mathematical model.

β is defined as the mathematical model of self-learning factor. Hot rolling process is mainly the use of addition and multiplication of self-learning. For the addition of self-learning are (see above)

$$Y = f(X_{1,}X_{2.}\cdots,X_m)+\beta \tag{3}$$

The study is about multiplication:

$$Y = \beta\times f(X_{1,}X_{2.}\cdots,X_m) \tag{4}$$

In the process of production, mathematical model output variable the actual value of Y and model of the input variables the actual value of X can be directly measured or indirectly calculated through the measurement instrument. The output process model output variables of the N times actual Y measurements, model of the input variables X actual measurements N times written, self-learning coefficient ", also called "measured instantaneous, remember to deduce the exponential smoothing method of learning. However, the instantaneous generally is not able to direct measurement. So, how to use the data ,to measure and calculate mathematical model of self-learning correction coefficient of measured instantaneous, is very important.

In addition, the calculation formulas for the instantaneous are given by

$$\beta_n^* = Y_n^* - f(X_{n1,}^*X_{n2,}^*\cdots,X_{nm}^*) \tag{5}$$

In multiplication, the calculation formula for the instantaneous are given by

$$\beta_n^* = \frac{Y_n^*}{f(X_{n1,}^*X_{n2,}^*\cdots,X_{nm}^*)} \tag{6}$$

Finally using exponential smoothing can get n + 1 "new" self-learning coefficient.

$$\beta_{n+1} = \beta_n + \alpha(\beta_n^* - \beta_n) \tag{7}$$

At the same time, in process control level of computer set the documents from the study of each the mathematical model. According to thickness, material conditions in different width divide the kind of record that d self-learning documents of steel. In rolling n block just, model setup program will withdraw from the self-learning documents to set value calculation. When n block steel rolling process in the learning process, according to the actual measurement data calculated from the study of the "values" β_n^*. Since the study coefficient "update".

Since the learning process model summarized as follows:

(1) The actual data processing. Including the collection of real-time data and the data processing.

Data acquisition. According to certain method, from basic automation computers in the actual data sampling data. Including data and the data with different projects, it is use kinds of item for learning.

Data processing. To choose good data processing. As the most tarsi, remove data, and then remove the minimum number of calculating average.

(2) Since learning condition of judgment. Detection of measured data, the rationality of the model test whether meet the conditions of learning. With real data set limit to check, judge whether the actual value deviation over a given limit, if the data, this piece of steel output alarm, no longer, to avoid measurement data of the error caused by abnormal and learning. At the same time, to human intervention too much as operators operating under the pressure of rolling mill position or too much, nor speed intervention, in order to avoid due to abnormal condition makes self-learning coefficient "bad".

(3) Self-learning coefficient update. That first calculated from each of the project "instantaneous values", then amended the instantaneous exponential smoothing, the last update the correction coefficient of self-learning. The new self-learning coefficient is stored into study document for the next documents, rolling.

Due to the characteristics of rolling mill in practical applications, it is use model of classification, the model of product information and coefficients adaptive mechanism. In order to enhance the real-time control system, through the short-term learning can get online adjustment and rolling schedule update; In addition, in order to enhance the adaptability of the control system, through long-term learning can make the rolling forecast process optimization of procedures and preservation effect good rolling data, update information coefficient. It can better solve the problem of static and dynamic model.

C. Self- learning of the flow stress model

In the process of rolling mill, the self-learning function of model is also involved in various aspects. Now, self- learning of the flow stress model is discussed through flow stress model for example.

(1) Model linearization

At present, widely application at least squares are discussed based on linear equations. And in the actual production and application of linear equation is often encountered, may be hyperbola and power function, the exponential function and so on. If the nonlinear regression, computation, so as to greatly to return to a linear model. For flow stress model:

$$f_s = c_1 e^{\frac{c_2}{RT}}\dot{\varepsilon}^{C_4} \tag{8}$$

Where f_s is flow stress; c_1 is hardness coefficient; R is pervasive gas constant; c_2 is activation energy coefficients; T is strip temperature; ε is the strain rate; c_4 is stress coefficient.

Convert to linear form:

$$\ln f_s = \ln c_1 + \frac{c_2}{RT} + c_4 \ln \varepsilon \tag{9}$$

Make $\ln f_s = y$, $\ln c_1 = Z$ can get linear equations

According to the (11) flow stress equation can write:

$$y = z + \frac{c_2}{RT} + c_4 \ln \varepsilon \tag{10}$$

This is the model of flow stress. And the three variables z, c_2, c_4 is worth to constantly revised undetermined parameter. Flow stress σ of actual measurement can be given by the rolling force equation as follows:

$$\sigma = \frac{F}{W \bullet Q \bullet \sqrt{R' d}} \tag{11}$$

Where F is rolling force; W is strip width, D is rolling; R' is the roller radius of deformation; Q is geometric coefficient.

Make $y_{meas} = \ln \sigma$, by the least squares:

$$E = \sum_{i=1}^{N} (y_{meas}{}^{i} - y_{\mathrm{mod}el}{}^{i})^2 \tag{12}$$

Where N is the effective ways of N number of times, $y_{\mathrm{mod}el}$ is $\ln f_s$ of (2) type.

It can be obtained by $\frac{\partial E}{\partial z} = 0$, $\frac{\partial E}{\partial c_2} = 0$, $\frac{\partial E}{\partial c_4} = 0$.

$$\sum_{i=1}^{N} z + (\sum_{i=1}^{N} \frac{1}{RT}) c_2 + (\sum_{i=1}^{N} \ln \varepsilon) c_4 = \sum_{i=1}^{N} (\ln \sigma^i)$$

$$\sum_{i=1}^{N} (\frac{1}{RT}) z + (\sum_{i=1}^{N} (\frac{1}{RT})^2) c_2 + (\sum_{i=1}^{N} \ln \varepsilon \frac{1}{RT}) c_4 = \sum_{i=1}^{N} (\ln \sigma^i \frac{1}{RT})$$

$$\sum_{i=1}^{N} (\ln \varepsilon) z + (\sum_{i=1}^{N} \ln \varepsilon \frac{1}{RT}) c_2 + (\sum_{i=1}^{N} (\ln \varepsilon)^2 c_4 = \sum_{i=1}^{N} (\ln \sigma^i \ln \varepsilon)$$

Where T and ε is relevant is to i in measured value and the calculated value. The matrix form into are as follows:

$$\begin{bmatrix} \sum_{i=1}^{N} 1 & \sum_{i=1}^{N} \frac{1}{RT} & \sum_{i=1}^{N} \ln \varepsilon \\ \sum_{i=1}^{N} \frac{1}{RT} & \sum_{i=1}^{N} (\frac{1}{RT})^2 & \sum_{i=1}^{N} (\ln \varepsilon \frac{1}{RT}) \\ \sum_{i=1}^{N} \ln \varepsilon & \sum_{i=1}^{N} (\ln \varepsilon \frac{1}{RT}) & \sum_{i=1}^{N} (\ln \varepsilon)^2 \end{bmatrix} \begin{bmatrix} z \\ c_2 \\ c_4 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} \ln \sigma^i \\ \sum_{i=1}^{N} (\ln \sigma^i \frac{1}{RT}) \\ \sum_{i=1}^{N} (\ln \sigma^i \ln \varepsilon) \end{bmatrix}$$

z, c_2, c_4 can be work out by using the Gaussian elimination method which are undetermined coefficient of three models.

(2) Optimization algorithm of self-learning

Through the three parameters: z, c_2, c_4 can be concluded the model function of flow stress to predict rolling force under a times[7]. Obviously, this way is calculated method of rolling piece of undetermined parameter model of slab of data. The function which has given, the poor reliability generous, cannot satisfy the requirement of process control. Only effectively use multiple data to meet the requirements of the exponential smoothing can achieve accuracy. In order to effectively by this method, it is usually use on the coefficient matrix exponential smoothing, make:

$$A = \begin{bmatrix} \sum_{i=1}^{N} 1 & \sum_{i=1}^{N} \frac{1}{RT} & \sum_{i=1}^{N} \ln \varepsilon \\ \sum_{i=1}^{N} \frac{1}{RT} & \sum_{i=1}^{N} (\frac{1}{RT})^2 & \sum_{i=1}^{N} (\ln \varepsilon \frac{1}{RT}) \\ \sum_{i=1}^{N} \ln \varepsilon & \sum_{i=1}^{N} (\ln \varepsilon \frac{1}{RT}) & \sum_{i=1}^{N} (\ln \varepsilon)^2 \end{bmatrix}$$

$$Y = \begin{bmatrix} \sum_{i=1}^{N} \ln \sigma^i \\ \sum_{i=1}^{N} (\ln \sigma^i \frac{1}{RT}) \\ \sum_{i=1}^{N} (\ln \sigma^i \ln \varepsilon) \end{bmatrix}$$

To:

$$A(1,1) = \sum_{i=1}^{N} 1, \ A(1,2) = \sum_{I=1}^{N} \frac{1}{RT}, \ldots, \ A(3,3) = \sum_{i=1}^{N} (\ln \varepsilon)^2$$

$$Y(1) = \sum_{i=1}^{N} \ln \sigma^i, \ Y(2) = \sum_{i=1}^{N} (\ln \sigma^i \frac{1}{RT}),$$

$$Y(3) = \sum_{i=1}^{N} (\ln \sigma^i \ln \varepsilon)$$

The coefficients are analysised with exponential smoothing, with A (1,2) as an example to explain:

$$\hat{A}(1,2)_{n+1} = aA(1,2)_n + (1-a)\hat{A}(1,2)_n$$

$\hat{A}(1,2)_{n+1}$ is forecast values of the first N + 1 times set A(1,2), $\hat{A}(1,2)_n$ is the forecast value of NTH set A (1,2),

$A(1,2)_n$ is measured value of N times control A (1,2).At the moment, it is calculated by the least-square method; α is the gain coefficient $0 \leq \alpha \leq 1$。It is get through the study of the number of points to effectively. Through the derivation:

$$\hat{A}(1,2)_{n+1} = \alpha A(1,2)_n + \alpha (1-\alpha) A(1,2)_{n-1} + \ldots + \alpha (1-\alpha)^n A(1,2)_1 + \alpha (1-\alpha)^n \hat{A}(1,2)_1$$

Because of a < 1, the farther from N + 1 times,the smaller using the data of the function, so that the continuity and effectiveness of the data are ensured.

III. CONCLUSION

This paper is based on the classical theory of rolling basis. Fluid stress model is put forward learning optimization control method of based on the exponential smoothing. It laid solid foundation for the mill control model adaptive optimization.

REFERENCES

[1] N.Zhang, Intelligent control theory and technology, tsinghua university, Beijing university press, 1997.
[2] D. Zhang, LiMou wiand SunYiKang, based on genetic algorithm, the cold rolling mill parameter optimization design system, Shanghai metal 2000,2222 (6) : 25-30.
[3] X. Cai, optimization and optimal control, Beijing, tsinghua university press, 1982.
[4] G. Wang, teaching model optimization, guangzhou: the cases of south China university of science and technology press, 1998.
[5] G. Shi, optimization method, Beijing,higher education press, 1999.
[6] J. Xue, the optimization principle and method, Beijing, The metallurgical industry press, 1983.
[7]X. Wang, Optimize Control of Model Adaptive for Two-Stand Tandem Reversing Steckel Mill, Ph. D. thesis, Kunming University of Science and Technology, Kunming, 2009

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 151-155

A Practical GPU Based KNN Algorithm

Quansheng Kuang, and Lei Zhao*
School of Computer Science and Technology, Soochow University, Suzhou 215006, China
Email: kqs.net@163.com, zhaol@suda.edu.cn

***Abstract*—The KNN algorithm is a widely applied method for classification in machine learning and pattern recognition. However, we can't be able to get a satisfactory performance in many applications, as the KNN algorithm has a high computational complexity. Recent developments in programmable, highly paralleled Graphics Processing Units (GPU) have opened a new era of parallel computing which deliver tremendous computational horsepower in a single chip. In this paper, we describe a practical GPU based K Nearest Neighbor (KNN) algorithm implemented by CUDA. In our algorithm, a data segmentation method has introduced in the distances computation step to adapt to the CUDA thread model and memory hierarchy. We obtain highly increase in performance compared to ordinary CPU version.**

***Index Terms*—K Nearest Neighbor, Data Segmentation, GPU, CUDA**

I. INTRODUCTION

For the past decade, the programmable Graphic Processing Units (GPU) has evolved into a kind of many-core processor with highly paralleled and multithreaded features. Compared with generic x86 based CPU, the current GPU provide tremendous computational horsepower and higher memory bandwidth. Nowadays, the GPU has been at the leading edge of chip-level parallelism and expanded the scope of application from 3D rendering to general purpose computing.

The KNN algorithm is a widely applied method for classification or regression in pattern recognition and machine learning. As a lazy learning, KNN algorithm is instance-based and used in many applications in the field of statistical pattern recognition, data mining, image processing and many others. The KNN algorithm is simple but computationally intensive. When the size of train data set and test data set are both very large, the execution time may be the bottleneck of the application.

In this paper, we propose a novel parallel KNN algorithm based on GPU. Our algorithm is specially designed for NVIDIA Compute Unified Device Architecture (CUDA), adopting the thread model and memory hierarchy of NVIDIA's GPU. A data segmentation method and a parallel Radix Sort are proposed to make full use of the computational horsepower of the GPU. As the results, on an inexpensive graphics card we can archive over 30X speedup than an ordinary CPU version. Therefore, KNN algorithm under huge number dataset and high dimension dataset are now practical and feasible.

The organization of the paper is as follows. Section 2 describes related work, including KNN algorithm and the programming architecture of the GPU. Section 3 presents the details of implementation of KNN algorithm based on GPU. In Section 4, experimental data are given and we conduct the analysis of the results. Finally, we conclude the paper in Section 5.

II. RELETED WORK

A. Principle of KNN algorithm

KNN algorithm is widely applied in pattern recognition and data mining for classification, which is famous for its simplicity and low error rate.

The principle of the algorithm is that, if majority of the k most similar samples to a query point q_i in the feature space belong to a certain category, then a verdict can be made that the query point q_i fall in this category. Similarity can be measured by the distance in the feature space, so this algorithm is called K Nearest Neighbor algorithm. A train data set with accurate classification labels should be known at the beginning of the algorithm. Then for a query data q_i, whose label is not known and which is presented by a vector in the feature space, calculate the distances between it and every point in the train data set. After sorting the results of distances calculation, decision of the class label of the test point q_i can be made according to the label of the k nearest points in the train data set.

Each point in d-dimensional space can be expressed as a d-vector of coordinates, such as:

$$p = (p_1, p_2, ..., p_n) . \quad (1)$$

The distance between two points in the multi-dimensional feature space can be defined in many ways. Using Euclidean distance is usually to be the most ordinary method, that is:

$$dist(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} . \quad (2)$$

Alternatively, Manhattan distance can also be used as:

$$dist'(p,q) = \sum_{i=1}^{n} | p_i - q_i | . \quad (3)$$

The quality of the train data set directly affects the classification results. At the same time, the choice of parameter K is also very important, for different K could result in different classification labels.

* Corresponding Author: zhaol@suda.edu.cn.

AP-PROC-CS-09CN005

The KNN algorithm is simple in calculation and can be applied to high-dimensional data sets. Nevertheless, when the test set, train set, and data dimension are larger than expected, the computational complexity will be huge and the operation time will be very long. When test set and train sets contain m and n vectors in d-dimensional feature space respectively, the time complexity of this algorithm is $O(m \cdot n \cdot d)$. At present, there are also some optimizations to improve the efficiency of algorithm, such as using KD-Tree to improve storage efficiency, or to lower precision for improve efficiency such as Approximate Nearest Neighbor Searching (ANN). There are also some papers present that some points in the train set take little or no effects to the final result, which could be cut to reduce the computational scale. In some cases, these methods can reduce the executing time by half.

B. GPGPU and CUDA

Nowadays, the theoretically performance of GPU is far more than that of GPU. The reason behind the discrepancy in floating-point computation capability between the CPU and the GPU is that the GPU is specialized for compute-intensive, highly paralleled computation, which exactly what graphics rendering does. Therefore, the GPU is designed to be more transistors in it are devoted to data processing rather than data caching and flow control.

Considering the huge computational horsepower delivered by GPU, methods were taken to make GPU play an active role in non-graphics purpose, which called General Purpose GPU (GPGPU). Nevertheless, applications specially designed for graphics hardware abstraction using graphics languages is difficult before CUDA appears. CUDA (Compute Unified Device Architecture), parallel programming model is designed to overcome this challenge by providing standard programming languages such as C to the programmers instead of imposing them to map non-graphics application through the graphics application programming interfaces.

Figure 1 shows the threads abstraction of CUDA. The host means the CPU while the device refers to the GPU.

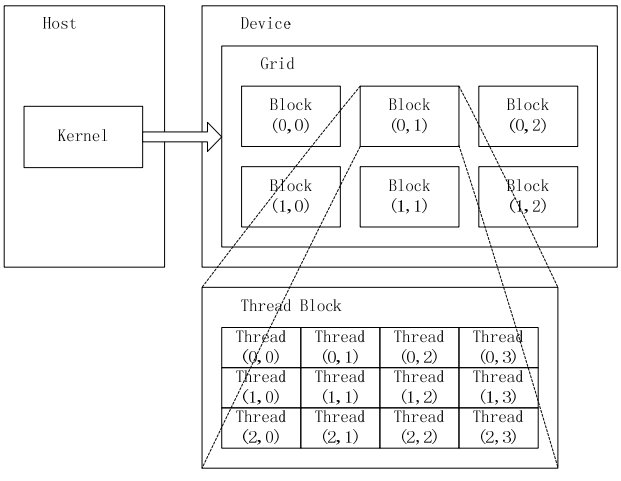


Figure 1. Threads Organization in CUDA

The beginning and the end of the application executed by the CPU must be serial code, in which one or more steps could be organized parallelism. The parallel code, which called "Kernel", is assigned to the device as a grid of Thread Blocks. The Thread Block containing hundreds of threads is dispensed to a Streaming Multi-processor (SM) for execution, which is composed by 8 Streaming Processors (SP). Each 32 threads in a Thread Block are organized into a Warp during executing. This is also referred as SIMT (Single Instruction, Multiple Threads) model.

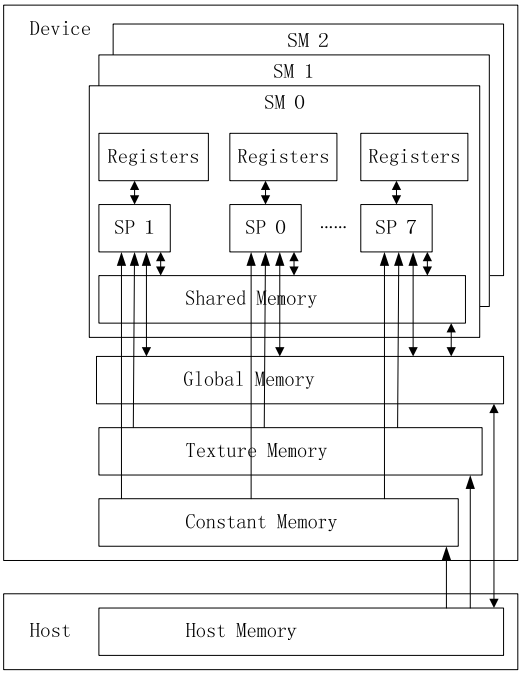


Figure 2. Memory Hierarchy in CUDA

Another important key point in CUDA architecture is the memory hierarchy. The register is the fastest but could only be accessible by a thread. Each SM contains 16KB of shared memory, which is shared by a Thread Block. The Global Memory is the video memory in the graphics card, which is usually have wide bandwidth and high frequency and much more faster than the host memory. The Texture Memory and the Constant Memory have the same speed as Global Memory but read-only and cached in the SM, as illustrated in Figure 2.

III. KNN ALGORITHM BASED ON CUDA

A. Overview

The basic process of KNN algorithm is as follows. First, the data pre-processing phase is to initialize the labeled d-dimensional train data set as well as the test data set to be classified. Second, select one test point in the test data set and calculate the distances between it and each point in the train data set. The next phase is to sort the results of distances computation, and find out K smallest results according to the parameter K. The fourth step is to determine the class label of the test point by the election result of K points. Finally, select another point in test data

set and go to step two repeatedly until the test data set is empty.

Euclidean distance is used in this paper. For the purposes of distances comparing, it is not necessary to compute the final square root in the Euclidean distance expression. So the squared distance will be used in phase two to reduce the computation.

According to the results of the analysis, the distances calculation phase can be highly paralleled and can reach a high speedup ratio in GPU implementation. The sorting step can also obtain benefits by using GPU acceleration. The remaining step, such as the determination of class labels are simple and consume little time that will be implemented on the CPU.

B. Segmentation method in distances computation

In the distances computation phase, distances between every point in the test set and each point in the train set should be calculated.

For the consideration to reduce program branching and to streamline operations, this paper adopts the way in matrices to restore multi-dimensional data set. The train data set A and the test data set B are both *d*-dimensional sets. That is, the number of features or columns to describe each vector in data set is *d*. The number of vectors or instances or samples in the train data set is *n*, while *m* for the test data set. Consequently, we restore the train data set as a $n \times d$ matrix in the memory, and a $m \times d$ matrix for the test data set. The result set C, which containing all the distances between each pair of points in A and B, is described in a $m \times n$ matrix. So the element in data set C which located in column *x* and row *y*, presents the distance between the vector in A whose row number is *x*, and the vector in B whose row number is *y*. As discussed earlier, the distances restored in C are squared Euclidean distances as the computation of square root does not affect the sorting results. The overall computational complexity of this phase is $O(m \bullet n \bullet d)$.

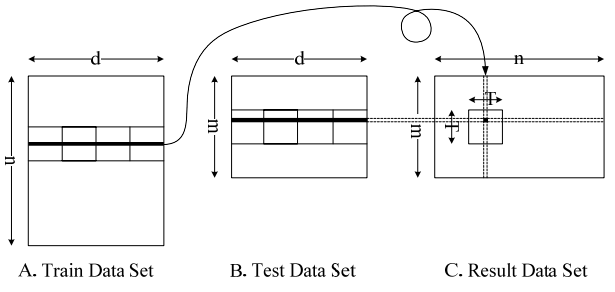


Figure 3. Data Segmentation Method in Distances Computation

The result data set C is divided into a large number of the tiles with the width of *T*. Each thread in the GPU takes charge of one element in C, i.e. computes one distance between a pair of vectors in A and B. Each Thread Block containing $T \times T$ threads that calculates one tile in C. Consequently, there are $(m/T) \times (n/T)$ Thread Blocks in all. In order to take full advantage of the high-speed Shared Memory in the GPU, we introduce a batch loading strategy when reading data from Global Memory. That is, each tile in A and the corresponding tile in B is loaded from Global Memory to Shared Memory for one step of calculation, when before each thread computes the square of difference between tow corresponding elements in A and B, and adds the result to the exact position in C. The batch loading strategy should be repeated d/T times to obtain a tile of squared Euclidean distances between *T* vectors in A and another *T* vectors in B. The pseudo code of the kernel function is shown as follows:

Algorithm 1: distances_computation(train, test, result)

```
{Each block is given the 2-Dimensional identifier bx, by, and tx,
ty for each thread. }
sub_result = 0;
temp = 0;
for each sub-tile in dimension/T do
    loads shared_train in train set to shared memory ;
    loads shared_test in test set to shared memory;
    syncthreads;
    for k=0 to T do
        temp = shared_test[ty][k] - shared_prob[tx][k];
        sub_result += temp * temp;
    end for
    syncthreads;
    add sub_result to the corresponding position in result set;
end for
```

Some limitations of CUDA specification are as follows: the maximum number of threads per block is 512; the maximum number of active threads per multiprocessor is 768; the maximum number of active blocks per multiprocessor is 8; the amount of shared memory available per multiprocessor is 16 KB organized into 16 banks, etc. According to the consideration of various constraints, it is appropriate to set T to the number of 16. Consequently, each Thread Block contains 256 threads, every Stream Multiprocessor would execute 3 Threads Blocks at the same time, and the number of Thread Blocks being parallel execution should be 3 times of the number of Stream Multiprocessor on the GPU.

This data segmentation strategy could make full use of the Shared Memory and could reduce reading and writing to the Global Memory. We can achieve 90X speedup in a low-end GPU than a CPU in the experimental result. Please refer to Section 4 for details.

C. Parallel Sort based on CUDA

Generally speaking, it is difficult to use GPU to accelerate sorting algorithms to a wondrous speedup ratio as in the distance computation phase, for there are too many branches in the thread and it's not fit for the GPU execution. Another reason is that the computational complexities of the current sorting algorithms are already very low. Among the CPU serial sorting algorithms, Quick Sort being the fastest one, is dominating the performance evaluation even in the same time complexity of $O(n \log n)$ algorithms when applying large amounts of

data. An implementation of a GPU-based parallel Bitonic Sort for huge data set introduced by us could bring a good performance of 10X~20X speedup compared to the CPU serial ordinary version of Bitonic Sort. However, it is not so significant compared to CPU Quick Sort.

Finally, the sorting algorithm we applied in this paper is a CUDA-based Radix Sort proposed in reference [7]. In a Radix Sort, it assumes that the keys are d-digital numbers and sorts one digit from least to most significant of the keys at a time. The implementation of the Radix Sort is divided into four steps:

1) Each block loads and sorts its tile in Shared Memory using b iterations of 1-bit split. Empirically, we can reach best overall performance by choosing b = 4.

2) Each block writes back the results to Global Memory, including its 2b-entry digit histogram and the sorted data tile

3) Conduct a prefix sum over the p×2b histogram table, which stored in column-major order, to compute global digit offsets.

4) Using prefix sum results, each block copies its elements to their corresponding output position.

This sorting algorithm can reach many times in performance compared with CPU Quick Sort. In the final performance test, the sorting phase occupied the largest proportion of the overall computing time. The sorting phase becomes the bottleneck in performance of the whole application.

D. Label decision

This step is to decide the classification label of the query point in test data set, according to the *K* nearest points in train data set. In this paper, a simple statistical election is made to complete the target among the labels of *K* points. As the result of the previous phase, we can get *K* nearest neighbor of a query point in the train data set, then we statistic the occurrences of each classification label. The most frequently occurred label would be chosen as the forecasting label of this query point. Weighted statistical methods can also be use in this step. The principle of this method is that the nearer neighbor to the query point should have a higher weight. We also have to define the weight values in advance in this method.

However, the computational complexity of this phase is low and it consumes little time. In our experimental results, no more than 20ms was spent during execution of this step. Meanwhile, the program branches are very high, and it is difficult to optimize for the GPU execution. Therefore, the CPU is adapted to accomplish this work.

IV. Experimental Results

A. Environments

The computer used to do this comparison is a Pentium D 2.8GHz dual core CPU with 1.5GB of DDR2 memory. The graphic card used is a G92 based NVIDIA GeForce 9600GSO with 96 streaming processors and 192bit 384MB of DDR3 memory interfaced with a PCI-Express 1.1 port.

The test data, Adult dataset, is from the UCI Machine Learning Repository. The number of classification label is 2, with 123 numbers of features. The values had been normalized into [0, 1] of real numbers. The a1a data set including a train set with the size of 30956 and the test data set is 1605, while the a2a data set including a train set of 30296 and the test set is 2265.

B. Performance and analysis

The CUDA implementation of GPU-based algorithm introduces by this paper is identified as “GPU”. For comparison, the original CPU serial algorithm is identified as “CPU” and Approximate Nearest Neighbor Searching using brute force method with KD-Tree optimization is identified as “ANN-Brute”. The initialization and input-output part of the program using in the methods are almost the same. The execution times of core part in each algorithm are shown in Figure 4.

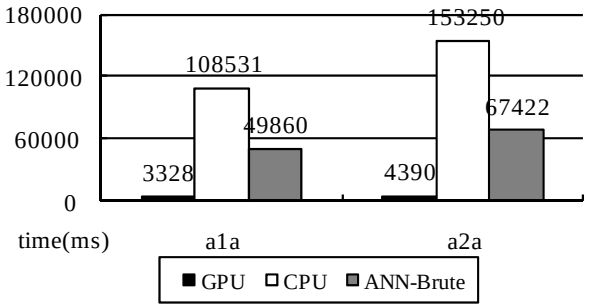


Figure 4. Execution Time-Overall

As is shown in the figure, the advantage of GPU algorithm in overall execution time is obvious. In a1a data set, we achieved the speedup of 32.61X compared with CPU algorithm and 14.98X with ANN-Brute method. In a2a data set, we reached the speedup of 34.91X compared with CPU algorithm and 15.36X with ANN-Brute method.

Since a kind of data segmentation method is presented in this paper and the tile size is 16, so we obtained appreciative performance in high-dimensional data sets. But if the data dimension is relatively small or even lower than 16, the performance will be reduced, when we should reconsider the tile size as parameter. In this condition, we can use rectangle tile instead of squared tile. For example, use an 8-width and 32-height tile, to adapt some small dimension data set, while the number of threads per block is still kept 256.

It could also be found during the experiments that, the decision of classification label phase in the algorithm is about 16ms. This phase is relatively simple and the execution time is very short and negligible, when Distances calculation and sorting phase occupying most of the time. Because the ANN-Brute algorithm as a whole process is completely different from GPU and CPU algorithms, we use Table 1 to illustrate the execution time in each phase in GPU-based algorithm and CPU comparison algorithm. The distances calculation phase is presented in T1, the sorting phase is T2 and the label

decision time is T3. The proportion of distances computation phase is shown in the last column.

TABLE I. EXECUTION TIME IN EACH PHASE

	T1 (ms)	T2 (ms)	T3 (ms)	T1/(T1+T2+T3) (%)
a1a GPU	828	2484	16	24.88%
a1a CPU	77871	30645	15	71.75%
a2a GPU	1141	3233	16	25.99%
a2a CPU	110492	42752	16	72.10%

As is shown in the table, the distances calculation phase is occupying the major proportion of the execution time in CPU implementation. In the CUDA implementation proposed in this paper, we had already achieved 94X and 96X speedup separately in this phase and made it occupying a smaller proportion of the overall execution time, transforming the bottleneck to the sorting phase. In fact, the complexity of sorting algorithm had already relatively low and most of the time was spent in reading and writing operation to the Graphics Memory, resulting that in the sorting phase we could only obtain 12X~13X speedup than that in CPU implementation.

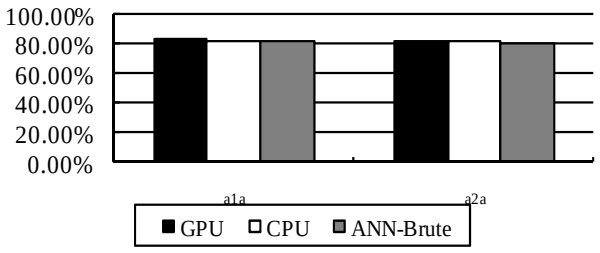


Figure 5. The Accuracy of Classification

Being a classification algorithm, the accuracy of the method should be presented as routine. Figure 5 gives us the accuracy of the algorithm. The accuracy of the three methods shows little difference, for the principle of them is just the same. In KNN Brute Search algorithms, the final result depends entirely on the quality of the data set. The slight difference of the results is due to sorting step- for the same distances between the query point and that from train set with different category label was cut by the parameter K differently because of unstable sorting algorithms.

V. CONCLUSION

This paper presented a CUDA based KNN algorithm, which could take full advantage of the computational horsepower of GPU and its multi-leveled memory architecture, making the performance of the method obtain greatly enhancement compared with CPU implementation. The tremendous increase in performance reached a cluster of computers, on which is only a PC with a 500RMB ($73) cost graphics card. This method is valuable for the KNN method in high dimensions, large amounts of data for applications.

ACKNOWLEDGMENT

The paper is supported by National Natural Science Foundation of China (No. 60873047) and Natural Science Foundation of Jiangsu Province of China (No. BK2008154).

The authors are grateful to all the people for helpful suggestions. The authors would like to thank all the reviewers for their helpful comments on earlier drafts of this paper.

REFERENCES

[1] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu. “An optimal algorithm for approximate nearest neighbor searching fixed dimensions”, Journal of the ACM, 45(6):891–923, 1998.

[2] D. M. Mount, S Arya. “ANN: A library for approximate nearest neighbor searching”, http://www.cs.umd.edu/~mount/ANN/

[3] Enhua Wu, Youquan Liu, “Emerging technology about GPGPU”, APCCAS. IEEE Asia Pacific Conference on Circuits and Systems, 2008.

[4] “NVIDIA CUDA Compute Unified Device Architecture: Programming Guide”, Version 2.3, July 2009.

[5] Feng Cao, Anthony K. H. Tung, and Aoying Zhou, “Scalable clustering using graphics processors”, Lecture Notes in Computer Science, Advances in Web-Age Information Management - 7th International Conference, WAIM 2006.

[6] Daniel Cederman and Philippas Tsigas, “A Practical Quicksort Algorithm for Graphics Processors”, In the Proceedings of the 16th Annual European Symposium on Algorithms (ESA 2008), Lecture Notes in Computer Science Vol.: 5193, pages 246 - 258, Springer-Verlag 2008.

[7] Nadathur Satish, Mark Harris, Michael Garland, “Designing efficient sorting algorithms for manycore GPUs”, Proc. 23rd IEEE Int’l Parallel & Distributed Processing Symposium, May 2009.

[8] Wenbin Fang, “Parallel Data Mining on Graphics Processors”, Technical Report HKUST-CS08-07, Oct 2008.

[9] V. Garcia, E. Debreuve, M. Barlaud.Fast, “K nearest neighbor search using GPU”, In Proceedings of the CVPR Workshop on Computer Vision on GPU, June 2008.

[10] Buck, Ian, “GPU Computing: Programming a Massively Parallel Processor”, CGO '07.

[11] Mark Harris, “Parallel Prefix Sum (Scan) with CUDA”, http://www.nvidia.com/object/cuda_home.html, 2008-1.

[12] Mark Harris, “Optimizing Parallel Reduction in CUDA”, http://www.nvidia.com/object/cuda_home.html, 2007-11.

[13] Xiaowen Chu, Kaiyong Zhao, Mea Wang, “Massively Parallel Network Coding on GPUs”, IPCCC 08.

[14] X.-W. Chu, K.-Y Zhao, and M. Wang, “Practical Random Linear Network Coding on GPUs”, Technical Report, Dec 2008.

[15] M. Suhail Rehman, Kishore Kothapalli, P. J. Narayanan, “Fast and Scalable List Ranking on the GPU”, 23rd International Conference on Supercomputing (ICS), June 2009.

[16] John Stratton, Sam Stone, Wen-mei Hwu, “MCUDA: An Efficient Implementation of CUDA Kernels on Multi-cores”, Technical report, IMPACT-08-01, March, 2008.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 156-161

The Empirical Study of Relationship between Enterprise Strategy and E-commerce

Yuantao Jiang, and Siqin Yu
School of Economics and Management, Shanghai Maritime University, Shanghai, China
Email: jiangytao@hotmail.com, ysq@shmtu.edu.cn

***Abstract*—E-commerce and e-business (henceforth referred to e-commerce) has moved to the forefront of technology priorities, and it can help firms to do everything faster, better and cheaper. More and more conventional brick and mortal firms regard e-commerce initiatives as offering strategic opportunities to transcend their normal operations. To take full advantage of e-commerce, firms need to look at themselves from an alternative perspective. The purpose of this study is to increase understanding of the relationship between corporate strategy and e-commerce, and to test the value of e-commerce on firm. By collecting the data, statistics method was used to test the hypotheses. A wide range of relations between strategy and e-commerce was analyzed to help firms develop e-commerce better. The results of the analysis show that there is a positive relationship between e-commerce and firm strategy.**

***Index Terms*—E-commerce, enterprise strategy, SCM, value chains, logistic**

I. Introduction

One should not ignore the fact that the value proposition of e-commerce includes the creation of new market opportunities through electronic channels. These electronically channeled market opportunities enable companies to lower transaction costs, reduce delivery times, improve customer services, and add convenience. Many companies have moved their business entirely to the web. Some are establishing subsidiaries, then spinning them off as separate online business entities. Others are investing in or merging with online startups. E-commerce is no longer an alternative, and it is an imperative. Many companies are struggling with the same problems: what is the influence of e-commerce on corporation strategy and how does e-commerce influence firm strategy?

II. Theory and Proposition

A. E-commerce strategy is part of corpoeration strategy

In common with any other business activity, e-commerce needs to be guided by corporation strategy. The implementation of e-commerce is the process by which an organization seeks to achieve its e-commerce objectives. Typically, the organization has a range of strategic options, which support the achievement of its objectives, such as reducing costs, increasing prices, streamlining operations, and so on. The key feature of corporation strategy is that it offers a clear statement of the basis for differentiation from competitors. The development and formulation for e-commerce has much in common with development and formulation for other business contexts or functions. There is a trend that on the road to developing e-commerce, firms may neglect the fundamentals, and overlook fundamental business principles and forget the integration with corporation strategy. E-commerce implementation must be aligned with whole strategy of corporation. The relationship between e-commerce and corporation strategy is dependent on whether the business is a pure-play or Internet start-up, or whether e-commerce is one of several channels through which the business delivers products and services. The extent to which e-commerce is integrated with other business strategies is also dependent upon the extent of integration of business activities. Some businesses have contained the perceived risk associated with e-commerce by creating separate companies for their e-commerce activities. Such a model inevitably leads to an independent e-business strategy. Therefore, we propose:

Proposition 1. There is a positive influence of corporate strategy on e-commerce. Companies that incorporate e-commerce with strategy will implement e-commerce effectively.

B. E-commerce can improve SCM performance

E-commerce does not just mean trading and shopping on the Internet. It means business efficiency in all supply chains. Some of large companies have implemented their Internet platform for Supply Chain Efficiency in the past years, and others of them will follow in the next few years. L.Y. Shen and Jana Hawley pointed out that e-commerce and SCM are complementary in nature and need to be studied together [1]. SCM is inherently information intensive, so e-commerce and SCM have an integral part to play in creating and facilitating new forms of SCM [2]. Nothing appears to have had the same effect on SCM as E-commerce, which resulted in changing the focus of SCM from engineering efficient manufacturing processes to the coordination of activities in the supply chain network through knowledge management [3]. In other words, E-commerce technically made the SCM viable and facilitated SCM use in different industries. There are many ways in which innovative information flows could be used within supply chains. So, we propose:

Proposition 2. Supply chains efficiency should be improved via e-commerce. Companies that increase their supply chains efficiency via e-commerce will outperform those that do not.

C. Strategy environment would make great impact on e-commerce development

In applying e-commerce, organisations need to have a good understanding on the types of models available for adoption. While there is no single unique classification system for the types of e-commerce models available, B2C e-commerce is generally less complex than any B2B solution, so a firm must start with B2C initiatives before creating B2B tasks. B2C and B2B are two separate concepts. For example, B2B e-commerce models are generally classified into four generic categories: merchant models; manufacturer models; the buy-side model; and brokerage models. Each of these models has different functional characteristics resulting in different models being more applicable or suitable to particular industries, markets or situations. B2C e-commerce refers primarily to the buying and selling activities over the Internet, including such transactions as placing orders, making payments, and tracking delivery of orders on the Internet. So, the focus of B2C e-commerce is typically on the customer side as well. All other stakeholders of the organization, including employees and suppliers, are generally not the main concern. It only relies on client-to-server or port-to-port data flow. B2B generally refers to the use of the web and Internet-related technology to connect the extended organization, including such entities as suppliers, employees, and regulatory authorities. Both of the two incorporate newly developed Web technology into organizational and business processes. The use of Web technology results in improved efficiency. So, Without the B2C infrastructure, it will be difficult for firms to incorporate B2B functions. Therefore, we propose that e-commerce matched with strategy environment would be a great benefit to the implementation of strategy.

Proposition 3. Those that match e-commerce deveopment strategy with business environment will bring positive impact on implementation of firm strategy.

D. E-commerce can optimize value chain system

But e-commerce is a disruptive innovation that initially tend to degrade performance but promise greater long-term potential [4]. The reasonable defination regards e-commerce as the use of computing and communicatioin technologies to engage in a wide range of activities up and down the value-added chain, both within and outside the organization.

The traditional impact of information technology on firm strategy is improving the efficiency and effectiveness of organizations. Rockart and Scott Morton have suggested that traditional information technology also can have important implications for the competitive position of the firm [5]. They employ a modification of Leavitt's organizational model, showing that information systems can affect competitive performance through their impact on strategic factors, such as management processes, personnel, and organizational structure. There are plentiful of published papers or techniques for identifying opportunities to support management processes with information technology. A number of authors have identified opportunities for the application of information technology to create competitive advantage. Two general approaches can be distinguished: a value-added chain analysis of the firm's operations and porter's framework for competitive analysis. In addition to the impact of information technology within an industry and its boundaries, it has more macroscopic effects as well, affecting the structure of different marketplaces. A firm may be able to improve its portfolio of industries by taking advantage of structural changes catalyzed by new technology. Alternatively, a firm can actively seek opportunities to exploit its technology-related skills and resources in new industries. This leads to Proposition 3:

Proposition 4. Optimizing value chains by e-commerce will bring positive impact on implementation of firm strategy. Companies that optimize their value chains by e-commerce will achieve their stategic objectives easily.

E. E-commerce can improve Logistic Performances

On one hand, logistics is the backbone of e-commerce. E-commerce that is not supported by Logistics cannot guarantee that customers get the right products or services in the right place and the right time. Logistics Efficiency means having the right product at the right place at the right time. On the other hand, E-commerce could increase an organization's ability to sense and respond to the market needs by collecting and disseminating market information throughout the organization. By these information, the organization could accurately assess or stimulate market demand and search for new markets. Therefore, corporations that make e-commerce integrated with its Logistics would be more likely to leverage complementary assets and achieve efficiency and effectiveness benefits.

E-commerce, a very critical supporting component of logistics, allows logistics information to be shared by all logistics nodes in the supply chain, actually improving overall performance of the enterprise in delivering customer wants. It is usually proposed that passing information on the supply chain to all businesses in the supply chain via information communication technology will improve performance. In fact, recent research has shown, via the supply chain "Beer Game", that simply passing information on to logistics nodes can have a detrimental effect. This is due to the fact that, as well as having more information available schedulers need to know what to do with it.

In manufacture industry, some organizations have been taking huge steps toward building information links with their suppliers and customers to capitalize on the Logistics toward organizing business processes. Many issues have arisen regarding the adoption of e-commerce within the Logistics area. The benefits of utilization of e-commerce for Logistics are supported by many theories and empirical studies in manufacture industries, and organizations involved with the automobile, steel and computer area have worked hard to promote the idea of integrated and collaborated Logistics with the help of e-commerce technology. Over the past two decades, business in virtually every industry of the world economy

has benefited from or at least has been influenced by the technologies of e-commerce and Logistics. E-commerce will shape and transform Logistics in a fundamental and sustainable way. Therefore, we propose:

Proposition 5. There is a positive relationship between e-commerce and Logistics. Firms that realize the importance of e-commerce in logistics information will outperform those that do not.

III. MEASURES

A. Sample Choice and Data Collection

The data used for this investigation was provided by a survey of the 367 firms with e-commercee. Three hundred and sixty-seven questionnaires were mailed to sample firms. A total of 218 questionnaires were completed and returned to the researcher. Based on descriptions of the value-propositions of these firms, respondents could ostensibly be grouped into either product-manufacturing firms or service providers to the manufacturing industry.

B. Resources

In measuring the degree of the influence of e-commerce on corporations strategy and SCM, the degree that e-commerce corresponds with internal and external situation, and the degree of optimizing business value chains by e-commerce, as well as the degree of interrelationship between e-commerce and logistics, we used a five-point liket scale, ranging from highly unfavorable to highly favorable with a defined nertral anchor [6]. Highly favorable was numerically coded at 5.0, while the highly unfavorable anchor was coded as 1.0.

C. E-commerce

We created a latent variable for e-commerce, e-commerce application. Laten variables are hypothetical constructs that combine two or more observed varialbes. Specific variable intems for e-commerce application includes market and purchase. Specific variable items for market were: degree of internet in market, marketer skill in e-commerce, and for purchase: degree of internet in purchase, buyer skill in e-commerce. The scale for market demonstrated high internal validity with factor loadings at 0.75 or higher. Cronbach's alpha was 0.76. For purchase, factor loadings were at 0.70 or more, and Cronbach's alpha was 0.85. Each of these alpha levels indicated good internal validity and reliability of the measures comprising the latent variable.

D. Corporation strategy

The concept of strategy has been borrowed from the military and adapted for use in business. To analyze the corporation strategy adopted by the firm, we measured it by answering each of four questions: (1) the degree of fulfiling sale objective this year than last year; (2) how about the position in the same indrusty this year than last year; (3) the degree of completing profit this year than last year; (4) the degree of finishing innovation in product and service this year than last year.

E. SCM

According to prior research, SCM is the active management of supply chain activity to maximize customer value to achieve a sustainable competitive advantage. The organizations that make up the supply chain are linked together through physical flows and information flows. So, SCM was measured by examining two distinctive components: suppliers, customers. Specific varibles for suppliers include: supplier network skill, upper and downer suppliers network skill, and for customers: customers' communication with firm on internet, customer knowledge on firm on internet. Each observed variable was factor and reliability analyzed. For suppliers, factor loadings were 0.82 or higner, with a Cronbach's alpha for the variable of 0.90. For customers, factor loadings were 0.79 or higher with a Cronbach's alpha for the variable of 0.78. The above analysis indicate good internal validity and reliability of the measures comprising the obseved variable.

F. Strategy environment

Strategy environment include internal and external ones. we mainly care about the external ones, that is as same as the porter's five forces: threat of substitute products, the threat of established rivals, and the threat of new entrants, the bargaining power of supplies and the bargaining power of customers. So we think that strategy environment comprised of two distinctive aspects of external environment: competitor and substitute. Specific variable items for competitor included: the knowledge on competitor's e-commerce, the knowledge on other situation of competitor, and the frequency of studying competitor. For substiture, we examined the knowledge the knowledge on substiture's e-commerce, the knowledge on other situation of substiture, and the frequency of studying substiture. For competitor, factor loadings were a minimum of 0.87 or higher and Cronbach's alpha was 0.76, well above the minimum threshold set for reliability. For substitute, factor loadings were a minimum of 0.89 or higher with Cronbach's alpha of 0.68, indicating good internal validity and reliability.

G. Value chains

A value chain is a chain of activities. Product pass though all activities of the chain in order and at each activity the product gains some value. It includes primary activities and support activities. So, the value chains is measured by two factors: product manufacturing, and product design. Specific variable items for prudct manufacturing were: degree of applying e-commerce in manufacture process, the knowledge of manufacture staff on e-commerce, and for product design: degree of applying e-commerce in design process, the knowledge of design staff on e-commerce. For product manufacturing, the measure was factor and reliability analyzed with minimum factor loadings of 0.83 and a Cronbach's alpha of 0.90. For product design, the factor loadings 0.84 and Cronbach's alpha 0.73 were higher than the threshold.

H. Logistics

Logistics is the management of the flow of goods, infromation and other resources, including people and energy, between the point of origin and the point of consumption in order to meet the requirements of consumers. Logistics management is that part of the supply chain which plans, implements and controls the efficient, effective forward and reverse flow and storage of goods, services and related information between the point of origin and the point of consumption in order to meet customers' requirements. So, logistics was measured by transportation, inventory, distribution. For transportation, we examined the degree of e-commerce application in transportation, the knowledge of transportation staff on e-commerce. And specific variables items for inventory include: the degree of e-commerce application in inventory, the knowledge of inventory staff on e-commerce, and for distribution: the degree of e-commerce application in distribution, the knowledge of distribution staff on e-commerce. Each observed variable was factor and reliability analyzed. For transportation, the measure was factor and reliability analyzed with minimum factor loadings of 0.77 and a Cronbach's alpha of 0.78. For inventory, the factor loadings 0.75 and Cronbach's alpha 0.85 were higher than the threshold. For distribution, the factor loadings 0.81 and Cronbach's alpha 0.87 were higher than the threshold. Table I presents the correlation matrix for the observed independent variables and their mean, standard deviation, and reliability alphas.

IV. Result Analysis

Due to large sample size, the normal distribution of the data, and the random selection of the firms used, the correlation matrix, as well as structural equation model, and path coefficients and critical ratios analysis method, were chosen to test the propositions. Table 2 presents the correlation matrix for the observed independent variables and their mean, standard deviation. To best capture the theoretical interdependencies between e-commerce and strategies, we analyzed the data using special structural equation modeling (AMOS5.0). This procedure allows for simultaneous analysis of more than one dependent variable. To explore the relationship between e-commerce and corporate strategy, we measured a latent variable for e-commerce application, based on the two observed variables, market and purchase, for the latent variable SCM based on two observed variables, supplier and customers, for the latent variable strategy environment based on two observed variables, compititiors, substitutes, and for the latent variable value chains based on two observed variables, manufacture and design process, and finally, for latent variable logistics based on three observed variables, transportation, inventory, distribution. Latent variables, a hypothetical constructs, consist of two or more observed variables, so, indicators that measure a latent variable should exhibit convergent validity, indicated by their correlation. For corporate strategy, we used only one of four observed variables, that are objective, position, profit and innovation.

We used multiple fit criteria to rule out measuring biases inherent in the various methods to insure that the model fits the data well. Table II shows the multiple fit statistics for the model. The Chi-square divided by the degrees of freedom was 0.59, which is under the suggested ratio of two, for the hypothesized indirect model, and the p-value was 0.61, which is greater than the suggested 0.05. The model's adjusted goodness of fit was 0.95, indicating a good fit with the data. The normed fit index(NFI) was 0.93, well above the 0.90 level that is considered to be acceptable. The root-mean-square residual was a very acceptable 0.06 for the indirect model, indicating a low difference between the observed and model-implied covariances. Hotelling's critical N was 218, well over the 200 mark considered acceptable, thus indicating that the data fit very will with the model. These fit indices indicate that the model is an accurate representation of the data.

To test these hypotheses, we examined the individual path coefficients and the critical ratios in the model. Table III and table IV shows the path coefficients and the critical ratios for the independent variables in the model. Fig. 1 presents the model. In proposition 1, we thought that there is a positive and significant relationship of e-commerce and strategy. The critical ratio for this path is 2.21, indicating strong support for this proposition at $z \leqslant 0.01$ level. In proposition 2, we thought that there is a

TABLE I
RELIABILITY AND CORRELATION MATRIX

Scale	measure	Mean	SD	1	2	3	4	5	6	7	8	9	10	11
1	Scale	3.52	2.25	0.76										
2	Scale	3.57	2.37	0.1	0.85									
3	Scale	3.52	2.38	0.02	0.12	0.90								
4	Scale	4.01	3.19	0.36	0.02	0.02	0.78							
5	Scale	4.06	4.00	0.12	0.01	0.01	0.08	0.76						
6	Scale	3.87	2.57	0.11	0.12	0.01	0.15	0.08	0.68					
7	Scale	3.97	3.85	0.02	0.11	0.12	0.02	0.01	0.02	0.90				
8	Scale	4.18	4.21	0.08	0.02	0.20	0.01	0.02	0.02	0.01	0.73			
9	Scale	3.25	3.57	0.02	0.01	0.11	0.08	0.02	0.10	0.01	0.02	0.78		
10	Scale	3.86	3.86	0.01	0.58	0.62	0.02	0.02	0.11	0.11	0.02	0.18	0.85	
11	Scale	3.49	4.11	0.08	0.02	0.01	0.01	0.12	0.01	0.02	0.01	0.2	0.11	0.87

Note: 1 represents market, 2 does purchase, 3 does suppliers, 4 does customers, 5 does competitor, 6 does substitute, 7 does product manufacturing, 8 does product design, 9 does transportation, 10 does inventory, 11 does distribution.

TABLE II
STRUCTURAL EQUATION MODEX

Fit statistic	Model	Recommended value
Chi-square/degrees of freedom	0.59	≤2.0
P-value	0.61	≥0.05
Goodness of fit index(GFI)	0.96	≥0.90
Adjusted goodness of fit index(AGFI)	0.95	≥0.90
Normed fit index(NFI)	0.93	≥0.90
Root mean square residual	0.06	Low values (0 = perfect fit)
Hotellings critical N	218	≤200

TABLE III
PATH COEFFICIENTS AND CRITICAL RATIO OF VARIABLES

Latent variable	Observed variable	Path coefficient	Critical ratio
e-commerce application	Market	1.00	N/A
	Purchase	0.52	1.52[1]
SCM	Supplier	0.21	3.58[4]
	Customer	0.32	2.66[3]
E-commerce applying Strategy	Competitor	0.66	2.17[2]
	Substitute	0.15	1.97[2]
Optimizing value chains	Manufacture process	0.58	5.52[4]
	Design process	0.47	4.83[4]
Interrelationship between e-commerce and logistics	Transportation	0.69	2.26[2]
	Inventory	0.36	3.85[4]
	Distribution	0.18	4.23[4]

Notes: 1 z≤0.10; 2 z≤0.05; 3 z≤0.01; 4 z≤0.001

positive and significant impact of e-commerce on SCM. The critical ratio for this path is 2.35, indicating strong support for this proposition at $z \leqslant 0.005$ level. In proposition 3, we predicted that e-commerce development should been based on present strategy environment. This proposition was also supported, with the critical ratio at 3.18 and $z \leqslant 0.001$ level. For proposition 4, in which we hypothesized a positive and significant relationship between value chains and e-commerce implementation, we also found strong support with the critical ratio at 2.17 and a significant *z* score at the z ≤0.01 level. For last proposition, we predicted a positive interrelationship between the e-commerce and logistics. This proposition was also supported, with the critical ratio at 3.21 and $z \leqslant 0.001$ level.

V. DISCUSSION

The study's object is to seek the relationship between e-commerce and corporation strategy. By empirically analyzing the influnce of e-commerce on corporation strategy, results suggested that e-commerce could benefit corporation performance and the positive relationship between e-commerce application on value chains, SCM, logistics and corporation strategy is in evdience. the effect of the right e-commerce developing on corporation strategy also was significant. Findings not only offered empirical evidence to confirm the positive relationship between e-commerce and strategy, but also contributed to the growing literature on e-commerce and corporation strategy. This finding is interesting in light of the recent world financial crisis and suggests that e-commerce must be given priority in corporate strategy. So, to achieve the corporation strategy, it is advisable for firms to adopt e-commerce based on the whole comprehension of competitior and substitute situation, and to apply e-commerce to the right business oprationes, that are value chains, SCM and logistics. If e-commerce is adopted and integrated in corporation strategy, firms may garner the appropriate resources so that it could create the core capabilities to compete in the virtual market.

TABLE IV
PATH COEFFICIENTS AND CRITICAL RATIOS OF PROPOSITION MODEL

E-commerce to firm strategies	Path coefficient	Critical ratio
The influence of e-commerce on strategy	0.55	2.211[1]
The influence of e-commerce on SCM	0.68	2.352[2]
E-commerce applying strategy	0.23	3.183[4]
Optimizing value chains	0.53	2.171[1]
Interrelationship between e-commerce and SCM	0.26	3.213[4]

Notes: 1 z≤0.10; 2 z≤0.05; 3 z≤0.01; 4 z≤0.001

VI. CONCLUSION AND FUTURE STUDY

Most Internet "pure plays" companies could not find sustainable profitability simply by excelling in the management of technology, information, and the customer behavior. However, those that viewed e-commerce as a stand-alone appendage to their business would be less likely to succeed in these efforts. After the testiny for five propositions we found significant effects of e-commerce on strategy implementation. The findings from our study suggest that while firms face many challenges in crafting strategies, it is critical to think of the role of e-commerce. Firms must regard their e-commerce initiatives as an integral port of their strategic objectives.

While the sudy makes an significant empirical contribution by exploring the hidden nature of e-commerce and corporation strategy, certain limitations should be considered when interpreting the study. First, e-commerce as an innovation should evolves through different stages of its developing process. It is obvious that the perceptions of e-commerce charateristics change with its different stages. The study of foucing on one stage represnets only a cross-sectional picture of the

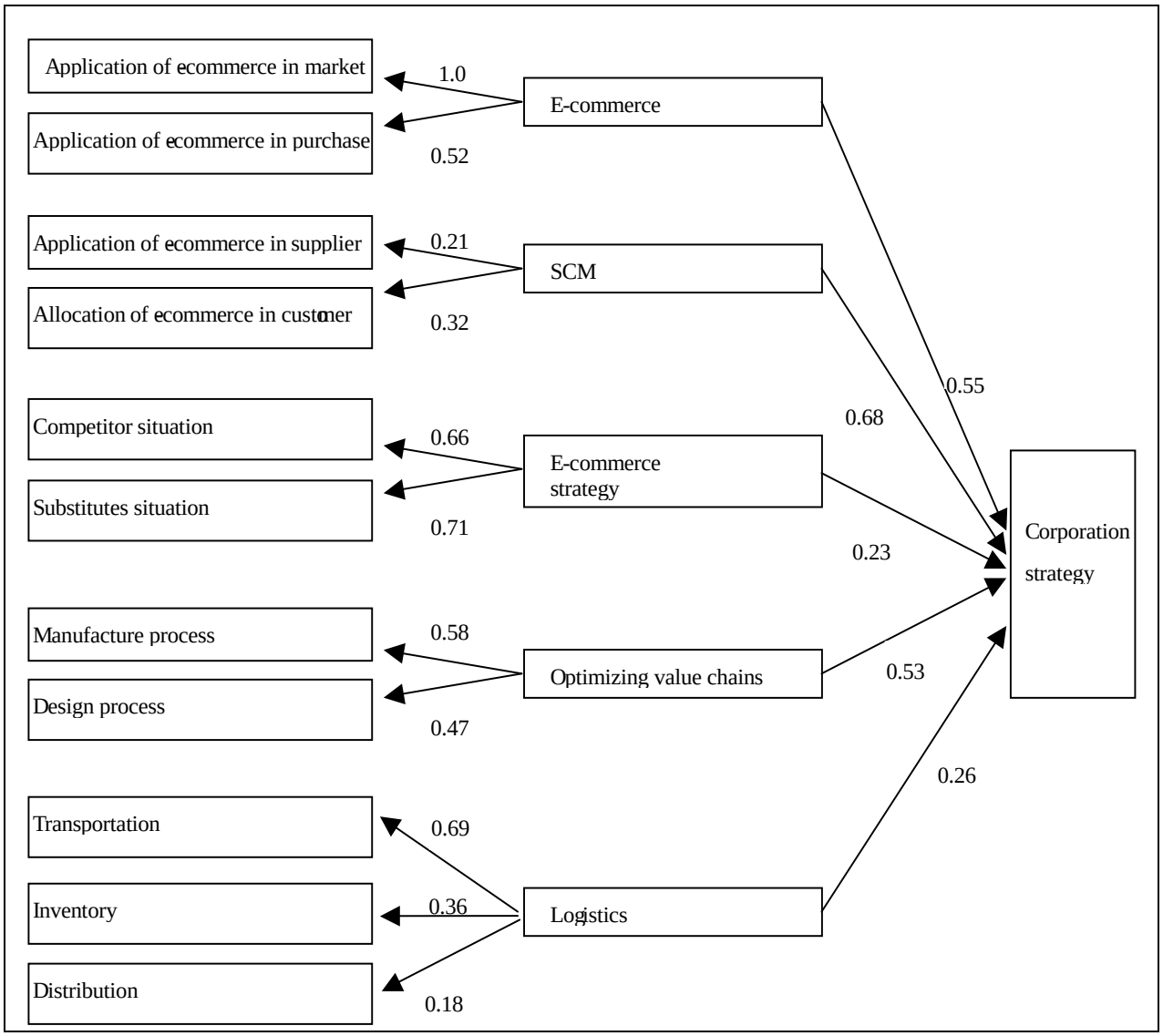


Figure 1. Path structure of model

relationsip between e-commerce and corporation strategy, so it just provides a snapshot of e-commerce adoption. Second, we looked at the relationship between e-commerce and corporation strategy, but we did not also control for the growth stage of the firm. It is possible that firms will have different strategy and e-commerce level at different lifecycle and that this will have different relationship between e-commerce and strategy. So, future research specifically examining e-commerce and strategy at different times in the organizational life-cycle, would address this issu.

ACKNOWLEDGMENT

This work was supported in part by a grant from Shanghai Key Subject Construction Program of Logistics and Distribution Engineering and Management(No.T0602) and the Important Program of Management Science & Engineering of Shanghai Maritime University (No. XR0101) and the Science & Technology Program of Shanghai Maritime University.

REFERENCES

[1] Liuying Shen, Jana Hawley, etc. “E-commerce Adoption for Aupply Chain Management in U.S. Apparel manufacturers”, Journal of Textile and apparel, technology and management, Vol.4, no.1, 2004, pp.1-10.

[2] Strader, T.J., Shaw, M. J. “Electronic Markets: Impact and implications”, in: Shaw, M., Blanning, R., Strader, T., Whinston, A. Handbook on Electronic Commerce, Springer, Berlin, 2000, pp.77-98.

[3] Menendez,T., and Achenbach, S. etc. Prenatal Rrecording of Fetal Heart Action with Magnetocardiography (in German), Zeitschrift für Kardiologie, No.2, 1998, pp.11-18.

[4] Lord C. “The practicalities of developing: a successful e-business strategy”, Journal of Business Strategy, Vol. 21, no.2, 2000, pp.40-43.

[5] Rockart, J. and Scott Morton, M.S. “Implications of changes in information technology for corporate strategy”. Interfaces, Vol.14, no.1, 1984, pp. 84-95.

[6] Linda F. Edelman, Candida G. Brush and Tatiana S. Manolova. “The impact of human and organizational resources on small firm strategy”, Journal of small business and enterprise development, Vol.9, no.3, 2000, pp. 236-244.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 162-165

The Design of Developed BP Arithmetic and Its Application in the License Plate Recognition

Meng Sun , Wenzheng Li , and Haisheng Li*
College of computer and information Engineering
Beijing Technology and Business University, Beijing 100037, P.R.China
*Corresponding author : lihsh@th.btbu.edu.cn

***Abstract*—The automatic vehicle license plate recognition (VLPR) is an important key technology in intelligent transportation system. License plate character recognition is a main step in this system. This paper presents an improved algorithm to accelerate the running speed of the network based on the analysis of traditional BP algorithm's running speed. Looking from the test effect, its recognition effect is good. Based on using the character classification recognition to distinguish, we will have the very big enhancement in again the efficiency.**

***Index Terms*—Automatic system of vehicle license-plate recognition, Neural network, Character recognizing**

I. INTRODUCTION

Due to a huge number of vehicles, modern cities need to establish effectively automatic systems for traffic management and scheduling. Automatic recognition of license plate is a very important part of intelligent traffic system and has been applied in many fields, such as the toll station, the intelligent management of residence zones and so on [2].

The key of an automatic system of vehicle license-plate recognition based on improved neural networks lies in extracting the characteristic of vehicle license-plate and the size of BP network. This paper presents an automatic system of vehicle license-plate recognition based on BP neural networks. Extract the feature of rough grid from characters of license plate, and withdrawing each picture element characteristic from normalized character as the input of neural network. In order to reduce the size of the network, encode the desired output of the output layer. Nodal point numbers of hidden layer and iterations reduced the size of BP neural network in the license plate character recognition system. Thus it ensures the accuracy and improves the recognition speed. Experimental results show that the correct rate of license-plate location is close to 100%, and the time of license-plate location is less than 1 second. Moreover, recognition rate of characters is improved due to improved neural networks and the known licenses-plate type. It is also observed that system is not sensitive to variations of weather, illumination, vehicle speed and the type of license-plate. In addition, it has the advantage of not requiring the information of the size of license-plate to be known in prior.

The Chinese standard's vehicles license plate is composed of seven characters. The first character is the Chinese character (abbreviations of the names of provinces), the second character is the capital letters of the alphabet, the third character is the capital letter of the alphabet or the digit, and other characters are the Arabic numerals. See Figure 1. See Figure 2. Vehicle license character recognition system training and recognition process mainly includes four parts. First, identify the characters to be normalized region (normalized into N × N pixel grayscale), and extract the normalized eigenvalue map. Then being the characteristic values as a neural network input, iteration repeated several times until the network training at all levels to achieve a stable weight, and preserved correspondence weight. Input the value of willing be to identify the Chinese characters, letters and numbers to neural network, and iteration repeated several times until the network training at all levels to achieve a stable weight. Finally, obtains the recognition result according to the node output's weight [3].

Figure 1. Standard Vehicle License

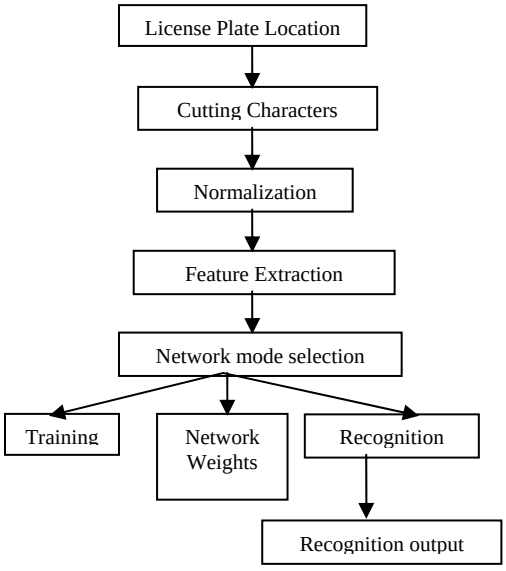


Figure 2. The process of neural network

II. Feature Extraction

The thick grid characteristic belongs to the statistical nature partial characteristic, and also known as gray-scale characteristics of the local. It is divided into N × N characters in a grid, and statistics for each of the pixel grid Su-volume. Each grid respective reflection character some part of characteristics. In the recognition stage, distinguishes the character by the character the statistical nature, which combined by each grid.

Firstly, the method of extraction thick grid feature is to be carried out to identify the size and location of characters normalized. Further divided into N×N-dimensional grids. Then statistics the number of the followed grid of black pixels (or white pixels). Thus obtains one by the numerical representation grid characteristic.

III. Neural Network Design

A. BP Neural Network

The BP network's learning algorithm has realized the Minsky's multi-layered network tentative plan [4]. It mainly used in function approximation, pattern recognition (with a specific output vector corresponding to the input vector), research fields, fuzzy clustering and so on. In the application of BP network, the most common is the hidden layer which contains only a three-tier network. This article will aim at the function of pattern recognition to use the BP network.

Error back-propagation (BP) algorithm is the most famous, most widely, but also multi-layer feed forward neural network algorithm of choice. The BP algorithm is composed of two parts, which are information forward transmission and error back-propagation. In the process of forward-propagating, the input information passes on from the input level after the concealment level cascade computation to the input level. Each neuron's condition only influence next neuron's condition. If has not obtained the expectation output in the output level, then calculated the change value of output level erroneous. Then turn to the anti-propagation, and instead passes on through the network the error signal along the original bridging fore-hearth revises each neuron's weight until to achieve the expectation the goal.

B. Improvement and Establishment of BP Neural Network

The BP network can study the massive pattern mapping relations, but does not need any knowledge of known mathematical functions to describe the relationship between their mapping. But the BP algorithm is one kind of gradient method essentially. If the search length of stride choice is inappropriate, the convergence rate will be very slow, and easy to fall into the partial minimum point. Therefore, many researchers proposed the improvement algorithm, which is mainly revolves four aspects to make the improvement [5, 6].

(1) To improve the adjustment method of learning rate parameter, such as the size of learning rate changes with the error gradient.
(2) Change the activation function, such as the S-type function to amend paragraph function composition.
(3) Weight correction method, such as the momentum of law, Newton's law.
(4) Improve the error function.

There are the progressive significance for improving the BP algorithm to convergence and overcoming local minimum with each kind of improvement algorithm. But there are also some limitations. Some effect of corrective method is not very obvious, and some methods are complex. It brings new question to algorithm restraining and computation dimension, which is difficult to achieve and more difficult to adjust parameters. In this article, I used the weight balance method, which performed the improvement to the concealment level's activation function. And combining this method to the additional momentum method [7], the auto-adapted study speed and the erroneous gradient's improvement, it has obtained the quite good effect.

IV. The Improvement of Hidden Layer Activation Function

The BP network's activation function generally have taken the S-type function at all levels, that is $f(x)=\frac{1}{1+e^{-x}}$, and its derivative is $f'(x)=f(x)[1-f(x)]$. But in the BP algorithm training process, we discovered that between the output level and the concealment level's weight adjustment quantity must be bigger than the concealment level input level by far the weight adjustment quantity. In the neural network training process, the weight adjustment quantity is too greatly easy to have the vibration and too small to making the convergence rate too slow. Because it has the obvious difference to the network training's contribution which is not only value weight adjustment quantity between the input level and the concealment level but also value weight adjustment quantity between the concealment level and the output level. When the value weight adjustment quantity is in the appropriate which is between the concealment level and the output level, the value weight adjustment quantity which is between the input level and the concealment level is too small, and it cannot participate in the learning process effectively. However when the value weight adjustment quantity is in the appropriate which is between the input level and the concealment level, the value weight adjustment quantity which is between the input level and the concealment level is too big, and it is easy to have had the accent. Therefore, in order to accelerate the network convergence rate, we should cause the two to be quite close. So that they could have the same level contribution and the coordination speeds up to the network training.

We have performed the improvement to the concealment level's activation function, that is changing $f_1(x)=\frac{1}{1+\exp(-\lambda_k x)}$ by original $f(x)=\frac{1}{1+e^{-x}}$, and its derivative is $f_1'(x)=\lambda_k f_1(x)[1-f_1(x)]$. Compared with the

original style, it add an adaptive factor λ_k , which will affect the form of S-function. When λ_k > 1, the gradient of the activation function will be increased and network convergence rate will be speed up. When λ_k <1, the gradient of the activation function will become smaller and the curve of the function will become smooth. λ_k can also balance the huge difference between weight adjustment quantity, which is between output level with intermediate level and input level with intermediate level. Taking λ_k >1, that is the weight adjustment quantity between the input level and the intermediate level is five times bigger than it, which between the outputs level and the intermediate level, the question that two weight adjustment quantity not balanced condition can be solved. Simultaneously because of the introduction of λ_k , it causes the curve of activation function changing steep, and also be easy to fall into the smooth area. The method used here is calculating separately.

A. Adding Momentum

Additional momentum causes the network when revises the weight not only to consider the error in the role of gradients, and to consider the influence in the erroneous surface of the impact of the trend, and its role is similar to a low-pass filter, which allows to neglect the small changes in network characteristics. Momentum in the absence of additional circumstances, the network may be the partial minimum , which can be over the minimum using the additional momentum's function. Improvement method based on the error back-propagation added on a direct proportion in the erroneous antipropagation's foundation in each weight's change with recently the weight change quantity value, and in accordance with back-propagation method to generate a new weight change. Therefore an improved algorithm for this purpose is proposed:

$$w(k+1) = w(k) + \alpha[(1-\eta)D(k) + \eta D(k-1)]$$

$w(k)$ could express not only a single weight but also the weight vector. $D(k) = \dfrac{\partial E}{\partial w(k)}$ is the negative gradient at the time of k. $D(k-1)$ is the negative gradient at the time of $k-1$. α is the learning rate and $\alpha > 0$. η is the momentum factor. The momentum item which is joined by the method of $0 \Leftarrow \eta \Leftarrow 1$ materially is equal to the damping item. It reduced the vibration tendency in learning process, thus improved the astringency.

B. Using Adaptive Learning Rate

An important reason for slow convergence with the standard BP algorithm is inappropriate choice of the learning rate. If we select a too small learning rate, the convergence rate will be too slow. If we select a too large learning rate, amendments are likely to be over and even to lead to oscillation or divergence. Hence we have a improvement measures to adjust the adaptive learning rate.

$$w(k+1) = w(k) + \alpha(k)D(k)$$

$$\alpha(k) = 2^{\lambda}\alpha(k-1)$$

$$\lambda = sign[D(k)D(k-1)]$$

The criterion of adjusting learning rate is to inspect whether the revised weight to reduce the error truly. If so, we can add a volume slightly which is larger than 1. If not, showing that it had been transferred, we should reduce the learning rate.

Ⅴ. BP NETWORK TRAINING

The collection of training samples must be comprehensive, and each training sample ought to be able to reflect the characteristic of every kind of recognition mode truthfully. In the license plate character recognition systems, the training sample of digital letter network is 10 numbers and 24 letters which is 0 to 9 and A to Z besides I and O, totally 34. The training sample of Chinese character network should include all 51 Chinese characters possibly which appear in the license plates. Because the character of license plate appears some phenomenon frequently, such as stroke adhesion, character displacement, we increase the character lattice to take the training sample suitably. This method can improve the stability of Chinese network performance and recognition rate. In view of improvement BP network, the study rate and the momentum factor should make the suitable adjustment along with the network restraining degree.

Network architecture and training parameter's concrete establishments of the two neural network in the license plate character recognition system as shown in Table 1.

And to ensure normal network convergence, two items which are the study parameter and the momentum factor should make the corresponding adjustment in the training process according to the erroneous change situation.

TABLE I.
CONCRETE ESTABLISHMENTS OF THE TWO NEURAL NETWORKS IN THE LICENSE PLATE CHARACTER RECOGNITION SYSTEM

	Alphabetic number neural network	Chinese character neural network
The numbers of nodes in input layer	512	512
The numbers of nodes in hidden layer	52	52
The numbers of nodes in output layer	6	6
Learning rate	0.1	0.1
Momentum factor	0.95	0.95
The value of erroneous target	0.01	0.01
The number of iteration	1542	1875

Ⅵ. ANALYSIS OF EXPERIMENTAL RESULTS

The simulation environment of this experimental is Matlab7.0. Non-linear transmission relations of nodes in network use Logsig which is S-transfer logarithm function. When using the classification of BP network, according to characteristics of the standard license plates,

we suppose that the output nerve number of the Chinese character and letter network is five, of the alphabetic number network is six, and of the digital network is four. We collected 100 images of license plates which are different in quality and size to analyze on the experiment, including Beijing, Lu, Su and other characters and letters, numbers, totally more than 600 character samples. Choose a comparatively standard sample from each kind of recognition sample to take the training sample, and all samples as recognition sample. Certainly, after these license plates having to a series of processing of binaryzation, localization division, correction for grade, character division and normalized and so on, we obtained the lattice as the input of neural network.

The key code is as follows:

```
createnn(P,T)  //create character recognition BP neural network and training function
function net = createnn(P,T)
alphabet = P;
targets = T;
net = newff(minmax(alphabet),[S1 S2],{'logsig''logsig'},'traingdx');  // create the BP neural network
net.LW{2,1} = net.LW{2,1}*0.01;  // level weight
net.b{2} = net.b{2}*0.01;  // bias vector
net.performFcn = 'sse';  // performance function
net.trainParam.goal = 0.01;  //error of training objectives
net.trainParam.show = 20;  // show interval steps of the results training
net.trainParam.epochs = 5000;  //the largest number of training steps
net.trainParam.mc = 0.95;  // momentum factor
P = alphabet;
T = targets;
[net,tr] = train(net,P,T);  //trains the BP neural network
```

The core code of the main program is as follows:

```
net = createnn(P,T);  // create the BP neural network by createnn(P,T)
[a,b]=max(sim(net,Ptest));  // start to identify
```

The training uses the traditional BP algorithm and the improved BP algorithm separately to carry on training the network. The error performance curve which is trained by traditional BP algorithm is shown in figure 3, and which is trained by the improved BP algorithm is shown in figure 4.

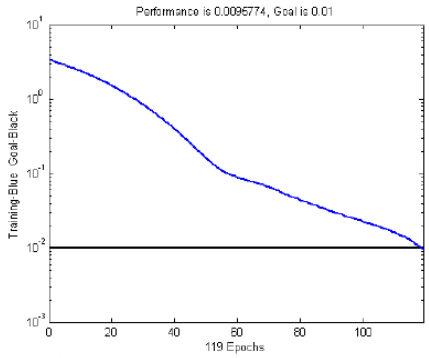


Figure 3. The error curve by traditional BP algorithm trained （node of the hidden layer: 9）

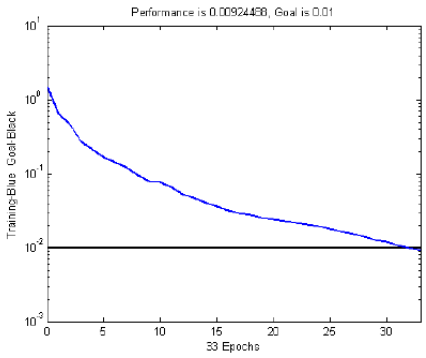


Figure 4. The error curve by modified BP algorithm trained（node of the hidden layer: 9）

Ⅶ. SUMMARY

Therefore, neural network with learning, anti-noise and parallel computing power license plate recognition is an effective method. This experiments proved that, antipropagation BP algorithm with the momentum item, the auto-adapted study rate training neural network in the character recognition with noise, has some characteristics which are the optimization overall importance and the accuracy, the quick convergence rate and so on. Looking from the test effect, its recognition effect is good. Based on using the character classification recognition to distinguish, we will have the very big enhancement in again the efficiency.

ACKNOWLEDGMENT

This research is partially supported by Science and Technology Development Program of Beijing Municipal Education Commission KM200910011007.

REFERENCES

[1] R. Parist, et al. Car Plate Recognition by Neural Networks and Image Processing. In Proc. IEEE International Symposium on Circuits and Systems,USA,2000,May 31-June 3.

[2] B. C. Coetzee, C. Botha, and D. Weber. Pc based number plate recognition system. Proceedings of IEEE International Conference on Industrial Eletronics, 1998.

[3] Liu Zhimin , Yang Jie , Shi Pengfei. Morphological detail algorithms [J] . Journal of Shanghai Jiaotong University , 1998 ,7 :32 - 39.

[4] Minsky M L, Papert S A. Perceptrons*: An Intro*duction to Computational Geometry , Cambridge：MIT Press.

[5] Engozinger S, Tomsen E. *An accelerated learning algorithm for multiplayer perceptions*: Optimization layer by layer [J]. IEEE Trans on Neural Networks, 1995, 6(1): 31-42.

[6] Ghellinck G. de and Vial J. P, *A polynomial Newton method for linear programming*. Algorithmica, 1986, 1(3): 425-453.

[7] Haodong Pei, Hongye Su, Jian Chu, The weight balanced algorithm of multi-layer forward neural network [J]. Acta Electronica Sinica, 2002, 30(1): 139-141.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 166-169

A Scheme of Rational Secret Sharing against Cheating

Yongquan Cai[1], and Huili Shi[2]
[1] College of Computer Science and Technology, Beijing University of Technology, Beijing, China
Email: cyq@bjut.edu.cn
[2] College of Computer Science and Technology, Beijing University of Technology, Beijing, China
Email: shihuili325@emails.bjut.edu.cn

***Abstract*—To address the issues and deficiencies of the traditional secret sharing scheme, we propose a rational secret sharing mechanism which applies game theory into secret sharing protocol. The paper mainly focuses on the secret sharing participants, who are rational, selfish and want to obtain the maximum benefit, how to detect cheat and take some punishable strategies to avoid being deceived under the condition of not revealing their secret share. By detailed analysis, we find that the scheme combining with the advantages of game theory and traditional secret sharing is feasible, secure and effective.**

***Index Terms*—secret sharing, game theory, repeated games, Nash equilibrium, trigger strategy**

I. INTRODUCTION

Based on the (n, m) threshold theory, secret sharing splits the secret into n shares. According to the convention, at least m or more participants can pool their shares and reconstruct the secret, but only m-1 or fewer shares cannot. This theory solves the problems of information storage, digital signature and so on. The secret sharing which was firstly proposed by Shamir [1] and Blakey [2] separately in 1979 has stimulated many scholars to research. With the wide applications of network, scholars focus on the security research about secret sharing, especially the secret sharing against cheating, and have presented many solutions for solving the problem [3~7]. Reference [3] first proposed the concept of verifiable secret sharing (VSS). Reference [4] gave a VSS scheme based on Shamir's scheme which can effectively prevent dealer from sending an error secret share but only check the correctness of his own share. Subsequently, reference [5] put forward a publicly VSS scheme which can verify the correctness of the shares sent by dealer, but can not guarantee that each participant submit the correct share. In addition, reference [6] proposed the first verifiable multi-secret sharing (VMSS) scheme. Besides, reference [7] presented a VMSS scheme without safe channel based on a two-variable one-way function. Although the dealer can't forge the secret share of all members to cheat in this scheme, it can give a publicly invalid function value enabling participants to restore invalid secret. So there are still deficiencies in these schemes.

Supported by Beijing Municipal Natural Science Foundation. (No.1102003).

Because of not getting rid of the limit of the classical algorithms, the previous schemes cannot effectively solve the cheats in secret sharing. This paper combines game theory [8~10] and classical theory [11] to solve the problem in the secret sharing.

II. THE RATIONAL SECRET SHARING SCHEME

A. Background Knowledge

Definition 1 Let G={N,S,u} represents a game, where $N = \{1, \ldots, n\}$ is the finite set of players, $S = \{S_1, \ldots, S_n\}$ is the list of strategies for all players and u={ $u_1, \ldots u_n$ } contains all the utility functions of each player. If executing G continuously, any player can choose next strategy based on previous game history. Then, we call the dynamic game as repeated games [12] and G is the stage game of the repeated games.

Let G (T) represents T times of finitely repeated games. The payoff function of player i can be represented as: $v_i = u_i(s^1) + \delta u_i(s^2) + \ldots\ldots + \delta^{T-1} u_i(s^T) = \frac{R_i}{1-\delta}$, where $u(s^t)$ is the Bernoulli payoff function, s^t is the strategy profile of G(t) ($T \geq t \geq 1$), δ is the discount factor, and R_i is the average discount payoff value of player i, where $R_i = (1-\delta)\sum_{t=1}^{T} \delta^{t-1} u_i(s^t)$.

Considering the repeated games G(t) of credit dilemma for n players, there are player 1 and player 2 which represents other players except player 1. For $i = \{1, 2\}$, let w_1 denotes the payoff of player i when i cheats and other players are honest; w_2 denotes the payoff of player i when i is honest and other players are also honest; w_3 denotes the payoff of player i when i cheats and other players also cheat; w_4 denotes the payoff of player i when i is honest and other players cheat. The relationship between them is $w_1 > w_2 > w_3 > w_4$. Table Ⅰ shows the game matrix of G.

Definition 2 If a vector of strategies $s^* = (s_1^*, s_2^*, \ldots, s_n^*)$ is a Nash equilibrium of

G={N,S,u}, we have $u_i(s_i^*, s_{-i}) \geq u_i(s_i, s_{-i})$, in which $i = 1, \ldots, n$ and $s_i \in S_i$.

TABLE I.
THE GAME MATRIX OF G

		Player 2	
		Honesty	Deception
Player 1	Honesty	w_2, w_2	w_4, w_1
	Deception	w_1, w_4	w_3, w_3

Definition 3 A trigger strategy is a class of strategies employed in a repeated non-cooperative game. A player using a trigger strategy initially cooperates but punishes the opponent if a certain level of defection is observed.

According to definition 2, we know that (deception, deception) is the Nash equilibrium of G in Table Ⅰ. In the case of considering discount factor δ and final payoff function, the final payoff of both rational players when they choose (honesty, honesty) in each stage may be greater than the value when they choose (deception, deception).

When not knowing the end time of the repeated games, player i's finitely repeated games can be analyzed as an infinitely repeated games. According to Table Ⅰ, the payoff function v_i is $\sum_{t=1}^{\infty} \delta^{t-1} w_2 = w_2 + \delta w_2 + \delta^2 w_2 + \ldots = \frac{w_2}{1-\delta}$ when player i takes an honest strategy in each game of G(T); when player i chooses deception in round r, other players will adopt trigger strategies in round r+1 while verifying player i' cheat, and the value $v_i^{'}$ of player i's payoff function is $\sum_{t=1}^{\infty} \delta^{t-1} u_i(G_t) = w_2 + \delta w_2 + \ldots + \delta^{r-2} w_2 + \delta^{r-1} w_1 + \delta^r w_3 + \delta^{r+1} w_3 + \ldots = \frac{(1-\delta^{r-1})}{1-\delta} w_2 + \delta^{r-1} w_1 + \frac{\delta^r}{1-\delta} w_3$.

And in order to make player i choose honesty rather than deception, and only if $\exists \delta$, for $\forall i \in \{1, 2\}$, it satisfies $v_i \geq v_i^{'}$ when $1 > \delta > 0$, $r \geq 1$ and $w_1 > w_2 > w_3 > w_4$.

Proof:

$$v_i \geq v_i^{'} \Rightarrow \frac{w_2}{1-\delta} \geq \frac{(1-\delta^{r-1})}{1-\delta} w_2 + \delta^{r-1} w_1 + \frac{\delta^r}{1-\delta} w_3$$

$$\Rightarrow (w_1 - w_3)\delta^r \geq (w_1 - w_2)\delta^{r-1}$$

$$\because 1 > \delta > 0,\ r \geq 1, w_1 > w_2 > w_3 > w_4$$

$$\Rightarrow \delta \geq \frac{w_1 - w_2}{w_1 - w_3} \quad (1)$$

So when δ meets (1), the final payoff function when every player choosing (honesty, honesty) is not less than of choosing (deception, deception).

B. Scheme Model

In a (n, m) rational secret sharing scheme, it supposes that m is the secret sharing threshold, n is the number of players, s is the sharing secret among all the players, $s_i (i = 1, \ldots, n)$ is the secret share. All the players are rational, selfish and their actions are designed to maximize the payoff function. Players exchange information at the same time and they are not aware of the end time of the repeated games. First, the dealer constructs the polynomials about secret reconstruction and share reconstruction, and then sends subshares to the players. Finally, the players reconstruct the secret by repeated games.

1) Scheme Initialization

The dealer chooses the discount factor δ which meets (1), two large prime numbers p and q, where q|p-1. g is the primitive element of Z_p, H(x, y) represents a two-variable one-way function which maps x and y to a fixed length. There is a notice board from which players can only read the information. And only the dealer has the right to add, delete and change the context of the notice board. The dealer publishes the value of (p, q, g, H).

2) Secret Distribution

The dealer can use the following steps to distribute these subshares among n players:

- Step 1 The dealer randomly chooses an integer r, $s_i^{'} \in Z_q (i = 1, \ldots, n)$ and computes the value of $s_i = H(r, s_i^{'})$. Then checks whether the value of $H(r, s_i^{'})$ is equal to that of $H(r, s_j^{'})$ when $j \neq i$ and $i, j = 1 \ldots n$ or not. If true, chooses the value of $s_i^{'}$ again. Otherwise sends s_i to player i through a safe channel and computes the value of $G_i = g^{H(r, s_i^{'})} \bmod p$. Finally, publishes the value of (r, G_i) in which $i = 1, \ldots, n$.
- Step 2 The dealer randomly chooses $b_l \in Z_q (l = 0, \ldots, m-1)$ and constructs (m-1)th degree polynomial $F(x)$ as follows: $F(x) = b_0 + b_1 x + \ldots + b_{m-1} x^{m-1} \in Z_q[x]$ where $s = F(0) = b_0$; computes $y_i = F(s_i) \bmod q (i = 1, \ldots n)$, then publishes $(y_1, \ldots, y_n)$.
- Step 3 The dealer randomly chooses $a_l^i \in Z_q (l = 1, \ldots, D_i)$ and constructs D_i th degree polynomial $f_i(x)$ where $|D_i - D_j| \leq 1$ when $j \neq i$ and $i, j = 1 \ldots n$, $d_i = D_i + 1$ and

$f_i(0) = s_i$ as follows: $f_i(x) = a_0^i + a_1^i x + \ldots + a_{D_i}^i x^{D_i} \in Z_q[x]$.

- Step 4 The dealer computes $s_{ij} = f_i(j) \bmod q$ for $i = 1, \ldots, n$ and $j = 1, \ldots, d_i$, secretly sends the set of $\{s_{i1}, s_{i2}, \ldots, s_{id_i}\}$ to player i, then publishes $g^{a_0^i}, g^{a_1^i}, \ldots, g^{a_{D_i}^i} \bmod p$. For example, constructs D_1 th degree polynomial $f_1(x) = a_0^1 + a_1^1 x + \ldots + a_{D_1}^1 x^{D_1} \in Z_q[x]$, where $f_1(0) = a_0^1 = s_1$ and $d_1 = D_1 + 1$; computes $s_{1j} = f_1(j) \bmod q$ for $j = 1, \ldots, d_1$, and secretly sends $\{s_{11}, s_{12}, \ldots, s_{1d_1}\}$ to player 1; publishes $g^{a_0^1}, g^{a_1^1}, \ldots, g^{a_{D_1}^1} \bmod p$.

3) Secret Reconstruction

- Step 1 Player i $(i = 1, \ldots, n)$ identifies whether the subshares sent by dealer are correct or not by (2). If true, player i continues the game according to the protocol. Otherwise quits the game.
- Step 2 In order to get the set of subshares of at least (m-1) other players, player i plays games with them and identifies the correctness of the subshares sent by other players by (2).

$$g^{s_{ij}} = g^{s_i} \prod_{k=1}^{D_i} (g^{a_k^i})^{j^k} \bmod p$$

$$(i = 1, \ldots, n,\ j = 1, \ldots, d_i) \qquad (2)$$

Proof:

$$g^{s_i} \prod_{k=1}^{D_i} (g^{a_k^i})^{j^k} = g^{s_i} \cdot g^{a_1^i \cdot j^1} \cdot g^{a_2^i \cdot j^2} \cdot \ldots \cdot g^{a_{D_i}^i \cdot j^{D_i}}$$

$$= g^{s_i + a_1^i \cdot j^1 + a_2^i \cdot j^2 + \ldots + a_{D_i}^i \cdot j^{D_i}}$$

$$= g^{s_{ij}}$$

- Step 3 Player i for $i = 1, \ldots, n$ computes the other players' secret shares by (3); for example, player 1 computes s_2 by $f_2(x) = \sum_{i=1}^{d_2} s_{2i} \prod_{j=1, j \neq i}^{d_2} \frac{x - x_j}{x_i - x_j} \bmod q$ where $s_2 = f_2(0) = a_0^2$

$$f_k(x) = \sum_{i=1}^{d_k} s_{ki} \prod_{j=1, j \neq i}^{d_k} \frac{x - x_j}{x_i - x_j} \bmod q$$

$$(k = 1, \ldots, n) \qquad (3)$$

- Step 4 Player i reads y_i for $i = 1, \ldots, n$ from notice board, and computes the secret s by (4);

$$F(x) = \sum_{i=1}^{m} y_i \prod_{j=1, j \neq i}^{m} \frac{x - s_j}{s_i - s_j} \bmod q \qquad (4)$$

In the repeated games, the processes between player i and player j for $j \neq i$ and $i, j = 1, \ldots, n$ are as follows: in the first round, player i sends s_{i1} to player j and player j sends his subshare s_{j1} to player i; in round r, player i chooses to send subshare deceptively to player j and j sends subshare honestly to player i; in round r+1, if player j verifies the subshare what i has sent in the previous r round is null, error, or illegal, j will adopt the trigger strategy and send nothing from the (r+1)th round to the end of the game.

None of the players knows when the repeated games will end. But they are aware of the fact that the degrees of the polynomials differ by at most 1. We analyze the three possible cases in round d of the protocol, where $d = \min(d_i, d_j)$, and $|d_i - d_j| \leq 1$ for $j \neq i$, $i, j = 1, \ldots n$:

- the number of player i's subshare is less than that of player j;
- the number of player i's subshare is equal to that of player j;
- the number of player j's subshare is less than that of player i;

Case 1: In round d, player i honestly sends his last subshare s_{id} to player j expecting that player j might have subshare $s_{j(d+1)}$. Otherwise, he might loose the chance of getting the subshare $s_{j(d+1)}$ and cannot recover the secret. In round d+1, as player i's subshares are exhausted without sending any subshare to player j. And player j does not know the subshares player i owning. If player j does not honestly send his subshare $s_{j(d+1)}$ to player i, he will loose the chance of obtaining the subshare $s_{i(d+2)}$ of player i. So player j will honestly send his subshare $s_{j(d+1)}$ to player i. After round d+1, both players can guess the number of each other's subshares, then they will compute the secret share to get the secret. Besides, we can analyze case 3 in the same way.

Case 2: in round d, in order to obtain other players' subshares, both player i and player j will choose honestly to send their subshares. In round d+1, their subshares are exhausted, so they send nothing. And when receiving nothing in the previous game, they can guess the number of each other's subshares. Thus, both players can reconstruct the secret.

In short, in the process of secret reconstruction, we conclude as follows: the discount factor δ which meets (1) gives rational players the incentive to carry out the protocol. If some player cheats, other players can check these deceptions without revealing their shares. Therefore, aiming to maximize the payoff function, rational players

will choose the honest strategy: send subshares to other players until the (d+1)th round to get the set of subshares of others' and finally reconstruct the secret successfully.

III. Security Analyses

A. The scheme has the advantages of traditional secret sharing in security.

When the number of any subset of all the players is less than m, they can't reconstruct the polynomial F(x) and can't get any information about secret s. Otherwise they can successfully recover the unique polynomial F(x) and the secret.

B. The scheme can dynamically add new player.

The dealer randomly chooses s'_{n+1} and computes the value of $s_{n+1} = H(r, s'_{n+1})$. If the dealer can't detect any conflict with others, he will send s_{n+1} to the new player and then publish $G_k = g^{H(r,s'_{n+1})} \bmod p$, construct polynomial of the subshare and send the set of the subshares to the new player.

C. The scheme can dynamically delete player.

The dealer can delete player k by deleting or changing the information, $g^{a_0^k}, g^{a_1^k}, \ldots, g^{a_{D_k}^k} \bmod p$, which is to verify the correctness of the subshare about player k. Without the verifying information, player k can't get through validation and can't obtain other players' subshares, so he can not recover the secret.

D. The scheme has fairness property.

The scheme can guarantee that rational players have no reason to deviate the protocol, and they will exchange information in a fair way. So that all players can either obtain information or get nothing from each other.

E. The scheme can check the deceptions.

It is impossible for illegal players to obtain the value of $H(r, s'_i)$ through the value of G_i because it means to solve the discrete logarithm problem. It is not feasible for illegal players to compute the value of s'_i when they know $H(r, s'_i)$ because of the unidirectional property of H(x,y), so they can not cheat other players by forgery. And in the process of the repeated games, illegal players can not forge by wrong subshare.

F. The scheme is anti-deceptive.

Although in the process of the repeated games, there exists some deceptions including player not sending message, sending error message and pretending legitimate player to cheat, other players can detect these cheats and will adopt some trigger strategies to minimize the playoff function of cheating player without revealing their own secret share. So in order to maximize the payoff function of themselves, all the players will correctly choose their strategies so that they can effectively prevent from being cheated. Therefore, this scheme is security.

IV. Conclusions

Applying the related knowledge of repeated games in game theory and classical secret sharing scheme, this paper constructs a new rational secret sharing scheme against cheating. Compared with the traditional secret sharing scheme, this scheme combines the payoff function with punitive strategy in game theory. Rational players prefer to maximize their payoff function values rather than deviating protocol to cheat. So even though some players cheat, other players can detect the cheat and will choose some trigger strategy without revealing the secret share to minimize the payoff of the players who cheat and finally effectively solve the cheat problem in the traditional secret sharing. Therefore, the scheme is a safe and effective.

References

[1] A. Shamir, "How to share a secret," Communications of ACM, 1979 ,vol 22(11), pp 612-613.

[2] Blakley G.R, "Safeguarding cryptographic keys," Proc.AFIPS 1979, National Computer Conference, 1979 , vol 48, pp 313-317.

[3] Chor B., Goldwasser S., Micali S. and Awerbuch B, "Verifiable Secret Sharing and Achieving Simultaneity in the Presence of Faults," Proc. 26th FOCS, 1985, pp 383-395.

[4] P.Feldman, "A practical scheme for non-interactive verifiable secret sharing," In Proc.28th IEEE Symp. on Foundations of Comp, Science(FOCS'87), IEEE Computer Society, 1987, pp 427-437.

[5] Stadler M., "Publicly verifiable secret sharing," In Advances in Cryptology-Eurocrypt'96, 1996, pp 191-199.

[6] Harn L, "Efficient sharing (broadcasting) of multiple secret," IEE Proc. Comput. Digit. Tech., 1995 , vol 142(3), pp 237-240.

[7] M.Hadian Dehkordi and Samaneh Mashhadi, "An Efficient Threshold Verifiable Multi-Secret Sharing," Computer Standards &Interfaces, 2008 , vol 30, pp 187-190.

[8] Maleka S., Amjed Shareef, C.Pandu Rangan, "The Deterministic Protocol for Rational Secret Sharing," IEEE, 2008.

[9] J.Halpern and V.Teague, "Rational secret sharing and multiparty computation: extended abstract," In 36t ACM Symposium on Theory of Computing(STOC), 2004, pp 623-632.

[10] Silvio Micali and abhi shelat, "Purely Rational Secret Sharing(Extended Abstract)," International Association for Cryptologic Research 2009, 2009, pp 54-71.

[11] JunShao, ZhenfuCao, "A new efficient (t,n) verifiable multi-secret sharing (VMSS) based on YCH scheme," Appl.Math.and Comput.2005 , vol 168(1), pp 135-140.

[12] Yao Guoqing, "Game Theory," Beijing, Higher Education Press, 2007.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 170-174

Effective Bandwidth Management Using Ajax Technology for E-Learning

Gaudence Uwamahoro[1], and Zuping Zhang [2]
[1] Central South University/School of Information Science and Engineering, Changsha, China
Email: first.gauwa2002@yahoo.fr
[2] Central South University/School of Information Science and Engineering, Changsha, China
Email:second.zpzhang@mail.csu.edu.cn

***Abstract*—New technique for web development is emerged as a powerful platform for building web applications with extensive client-side interactivity. The Ajax technology which is based on existing web technologies nowadays is used to build more responsive web applications. Ajax makes a significant advancement in a website. It delivers web contents better, smarter and richer. Ajax enhances the user's experience of application significantly with improvement of interactivity and speed in the application. The benefits of Ajax are adapted by E-learning to serve the application to be faster and responsive in order to help students to interact with the system, especially during the examination. With this new technology, amount of E-learning server load will be more decreased and traffic data between server and student computer will be reduced at maximum, which will help the student to get enough time to do the exam. Only small page bits are requested and sent to the browser as they are requested by the student, not the whole page at once, and as well small data will be transferred from the student's computer to the server which will reduce the E-learning web server load and so decreasing the bandwidth usage.**

***Index Terms*—Ajax, E-learning**

I. INTRODUCTION

Nowadays internet is used as source of knowledge in this world where people are intending to acquire knowledge as quick as possible using learning via internet. As reported in Ref. [5], Web based learning often called online learning or E-learning has become a new way in learning. E-learning is used by many universities in developing countries to promote the education of their people. Current E-learning technology has some drawbacks in its processing that can be obstacle for the student who is intended to get all satisfaction while he is learning.

The student needs E-learning to be faster and responsive for a good interaction. After finishing course, the student is evaluated from the exam passed. The examination process has to be considered with more importance because it is an evaluation tool for a student to know his understanding level of the course. Lack of enough bandwidth, can weaken the examination process. Reducing the data transfer between E-learning server and student's computer can decrease the efficient use of bandwidth and the server load problems.

In Ref. [6], the Ajax (Asynchronous JavaScript and XML) technology is used to solve those problems using the ability of sending and receiving necessary information and this will reduces web server load. Ajax is an emerging advanced technique used in web development to make web content faster and is a work-together of several pre-existing technologies. During the process of examination, the student does not have to undergo any kind of interruption. From Ref. [4], Ajax application avoids any interruption that can occur during interaction using Ajax engine as a new layer between web server and user to make application more responsive. The process the student uses in the examination has impact on results. If the student does the first question of the exam and if amount of data transferred between server and student's computer is minimized, that can make examination process faster.

We propose the development of E-learning web application using client-side caching and Ajax in order to minimize the number of client requests to the server. Better performance, more responsive interface and reduced or eliminated waiting time are the advantages that E-learning beneficiates using Ajax. Small amount of data transferred from the server as well as reduced number of client requests to the server is a result that shows a significantly high performance of Ajax application that also makes the Ajax technique special for data intensive application as well as for low bandwidth networks. The improved performance leads to much more responsive interfaces, which create the illusion that updates are happening instantly. As stated in Ref. [1], Ref. [3] and Ref. [6], in Ajax-based applications only the relevant page elements are updated with the rest of the page remaining unchanged and this approach eliminates the white screen and significantly decreases the idle waiting time.

II. RELATED WORK

Many researchers in software engineering have worked on web applications area, especially dynamic web applications. The traditional web applications like E-learning enables more bandwidth and server load where the server is required to load a full page for every request.

Supported by Hunan Province Natural Science Fund under Grant No. 07JJ6122 and Postdoctoral Science Fund of Central South University. Corresponding author, Zhang Zuping.

The emergence of advanced web technologies was the reason for the web applications developers tried to improve the traditional web applications using Ajax. A pertinent example has been the case of E-learning web applications. Many researchers on E-learning have used that technology for its improvement. Examination process is a very important part of E-learning that interested many researchers. From Ref. [3], the data transfer between the server and student computer can allow efficiently use of the bandwidth and reduce server load using Ajax to minimize traffic between student computer and server.

In examination process when many students are considered to start at the same time, the server loads the full page. At the second time when all students start for the first question, the next question is downloaded from the server. After finishing one question, the answer is sent to the server and so on. Downloading all questions at the same time takes a long time and sending each one response alone, will lead to the waste of bandwidth usage and server load which is not enough to make examination faster and convenient to the student and that is why we propose caching method to minimize the number of requests sent to the server and reducing time of request in order to make E-learning more efficient. This method shall solve the problem of lack of enough bandwidth and server load, considered as obstacle for student during the exam.

III. AJAX OVERVIEW

In this paper we proposed Ajax (Asynchronous JavaScript and XML), an emerged technology for web applications that helps to build more interactive web applications. This technology was coined by Jesse James Garrett in 2005. Ajax is not really a new technology because it is a combination of existing technologies which are already in use by traditional web applications. Ajax uses HTML (HyperText Markup Language) and CSS (Cascading Style Sheets) for data presentation, DOM (Document Object Model) for dynamic display and interaction, XML (Extensible Markup Language) and XSLT (Extensible StyleSheet Language Transformations) for data interchange and manipulation, XMLHttpRequest for data retrieval, and JavaScript which binds everything together. In the classic web applications, the communication is performed using directly http Request.

From literatures of Ref. [4] and Ref. [6], Ajax architecture differs from today's web applications architecture by adding a client-side engine called Ajax engine used as intermediate or new layer between user interface and server. The user activity leads to program calls to the client-side engine instead of a page request to the server and XML data transfer between server and the client-side engine. Moreover, Ajax engine is a collection

of JavaScript code that instantiates and uses XMLHttpRequest object to handle all communication with the server. As without this engine every user event would go back to the server for processing. All requests for data to the server are sent as JavaScript calls to this engine.

As reported in Ref. [9], Ajax engine requests information from the web server asynchronously. That means that the user can continue to perform other actions while the request is processed by the server, which is different from the synchronous system used in classic

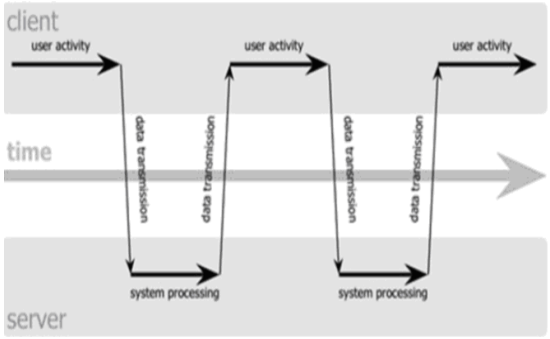


Figure1. A traditional web application is a synchronous system

web application which requires the user to wait for the server refreshing entire page in order to send him/her the first page of system before he/she can request the second which makes user operation interrupted as shown in Fig. 1 above.

Using Ajax, the page is loaded entirely only once the first time it is requested. After the page is loaded Ajax engine displays only the information needed by the user without reloading the entire page. Thus the waiting time is beneficiated and the web server load is reduced. As Ajax is asynchronous system as reported in Ref. [6] and Ref. [9], they noticed that when a user sends a request to the server, there will be neither waiting nor interruption for response. Because of that fact, the student after finishing a question he/she sends the response and starts the next question immediately. This produces the feeling that the information is displayed immediately as in Fig. 2.

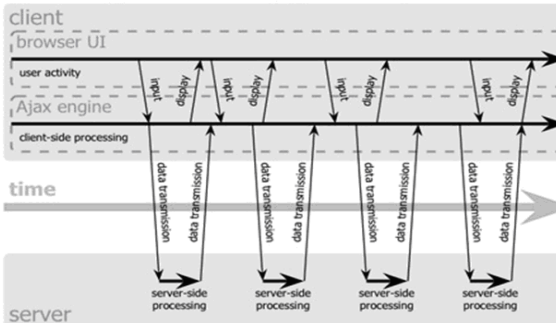


Figure 2. An Ajax web application is an asynchronous system

From Fig. 3, we were comparing classic web and Ajax technologies. We noticed that synchronous system can communicate directly with the server using http request whereas asynchronous system can communicate with a new layer between server and browser. In synchronous

system, when the user requests a page for the first time the server sends full HTML and CSS code at once ass is shown in Fig. 3.

That can happen when the student starts the exam and requests for the first question. After the page is loaded; if the user requests for the next question, the server processes the information, rebuilds the page and sends full page back to the browser Ref. [1] and that is shown in Fig. 3.

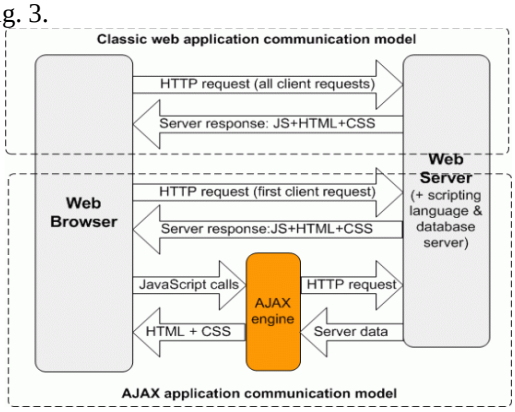


Figure 3. Comparison of classic web application model and Ajax web application model

In Ajax application communication model, the full page is loaded only one time when a user requests for a page for the first time as is in synchronous system or classic web application communication model. When the user makes a new request considered as next question, the Ajax requests only for small information to the server and displays the returned information without reloading the entire page. In this process JavaScript is used for asynchronous request and retrieves data from remote server and it is also used to extract data from XML. XML is used to collect numerical or text style data to the browser. Displaying of information returned from the server is done by HTML and CSS.

IV. Caching Method

We propose a new approach that uses client caching using JavaScript associative arrays as client cache to make e-learning faster and more responsive. Our proposed method comes to support improvement of examination process by minimizing the request bandwidth usage and time request. For the first time, the server loads full page once when student requests the first question. If the student starts the first question, Ajax engine is responsible for making further request of other questions and stores them into the client cache. A second cache is used to store the answers. If some questions are present in the answer cache, the Ajax engine would make further request with the transportation of an answer to provide efficient usage of the bandwidth. After finishing all questions, the student will submit the examination answers to the server, in this case only those answers which are remained in the answer cache will be sent to the server. The fig.4 above shows the interaction process of E-learning especially in examination using Ajax.

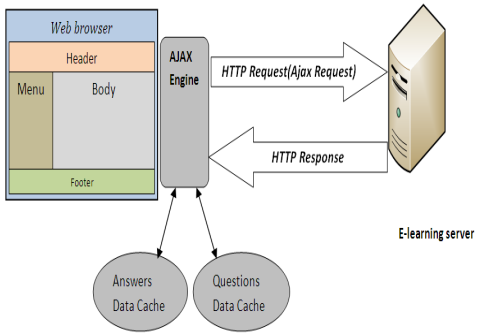


Figure 4. E-learning caching model

A. Reduce Server Load and Minimize Size of Response

The application consists of a typical page layout with a central section containing the changing content. The header, footer and navigation menu do not change during the application operation. Using traditional method, that content would have to be reloaded on every request. Ajax helps to reduce server load by not reloading a page where Ajax is used. However, using Ajax, a web application can request only the content that needs to be updated, thus drastically reducing bandwidth usage and load time. If the student starts the exam, E-learning server must load full page that contains header, navigation menu, footer and body where body contains the first question at the first time.

In the traditional method, when student requests the second question, the server loads again a page with header, navigation menu and footer and second question. That increases loading time and bandwidth misusing. But using Ajax, Ajax engine requests only the second question without reloading whole page. Fig. 5 illustrates an example.

Let 50KB be the size of header and *3.12s* (seconds) be the time used to load it. Let 40K be the size of navigation menu and *2.5s* is the time to load the navigation menu. The size of footer is 30KB and *1.85s* are needed to load it. The total size of header, navigation menu and footer is 120 KB and the size of each question is not fixed. The question one has 8KB and server use *0.5s* to load it, the question 2 has 6KB and is loaded using *0.37s* . For the first time the server loads entire page that has 128KB as size and *7.97s* for loading time. For the second question, the server loads 126KB and that takes *7.84s*. Using Ajax, instead of loading 120KB, the server loads only 6KB for the second question using 7.84s because that size is only data needed by student to do the question 2. The same process is used for question 3, question 4 and question 5 as is shown on Table I below.

The following abbreviations are used in Table I. Q represents question, HS represents header size, TL represents time to load, MS is used to represent menu size, FS for footer size, TT represent total time, TSL

TABLE I
MEASUREMENT OF AJAX WEB PAGE SIZE

Q	HS &TT		MS &TT		FS &TL		QS	QLT	TT	TSL
Q1	50	3.12	40	2.5	30	1.85	8	0.5	7.97	128
Q2	50	3.12	40	2.5	30	1.85	6	0.37	7.84	126
Q3	50	3.12	40	2.5	30	1.85	3	0.18	7.65	123
Q4	50	3.12	40	2.5	30	1.85	12	0.75	8.22	132
Q5	50	3.12	40	2.5	30	1.85	7	0.43	7.9	127

represents total size to be loaded, QS represents question size and QLT is used to represent question load time. We assume that we have used the internet connection of 128kbps to calculate loading time. Ajax eliminates the common contents that are header, navigation menu and footer that are reloaded on every request of question.

Thus, the loading time is decreased and a considerable amount of bandwidth usage could be saved. In the Table I above the unit of HS is KB, TT unit is s(seconds), MS unit is KB, TT unit is s, FS unit is KB, TL unit is s, QS unit is KB, QLT unit is s, TT unit is s and TSL unit is KB.

B. Minimize Number of Requests

If we use caching method it is possible to minimize number of requests to the server in order to efficiently use the bandwidth and decrease the data traffic between server and client. When a student finishes a question, Ajax engine sends it to the answer cache. Consider the student starts the exam; when the first question is processed, the Ajax engine requests next question to the server and server responds by sending the question to the Ajax engine. Ajax engine as intermediate is responsible to know if it is necessary to send the question to the student. When the student finishes, he/she shall send answer and requests the next question. Ajax engine sends the answer to the answers data cache and sends the next question requested by the student at the same time from the questions data cache. Sending the question from the answers data cache reduces time of request.

Ajax engine is responsible to know when to transmit answers to the server according to the number of responses in the answers data cache. If the answer has a big size, it can be transmitted alone, but the answers with small size are aggregated before being sent to the server. When data is transmitted from the client computer to server, there is some information called headers added to the data to allow moving on the network. Headers are different from layer to other. In Ref. [11], they noticed that when a layer receives a data it adds its own header before transmitting it to the next layer. In this paper we consider the headers used by the protocols in application layer, transport layer, network layer and data link layer. We use Http protocol with 566 bytes for header size. As stated in Ref. [10] and Ref. [12], the TCP (Transmission Control Protocol) header has fixed size of 20 bytes, the IP (Internet Protocol) header has 20 bytes and data link layer header has 22 bytes.

The size of data to send has impact on data transfer between E-learning server and the student's computer. To send answer one by one to the server increases the data traffic and misuses the bandwidth. The answers data cache helps to make aggregation of requests from the answers stored in. Aggregation of multiple requests through data caching method reduces the number of requests to the server which allows efficient use of bandwidth to make network efficient. As each request to the server requires its own overhead, aggregating multiple answers in a single unit of request reduces the number of requests and the bandwidth is used efficiently because the size of overheads needed for each request is decreased.

We are based on 5 questions prepared as exam. Let 30 bytes be the data size considered as the size of one question, 566 bytes for Http headers, 20 bytes for TCP header without any option, 20 bytes for IP header without any option and 22 bytes for header added by data link layer. Transmitting one answer needs 628 bytes of headers. The total size of one request is 658 bytes. Using classic web application technology which requires sending the answers to the server one by one without considering the size of request, sending every answer requires 658 bytes. Comparing to the Ajax technology and client caching method which allows aggregation of requests, it is possible to send one or more answers to the server in one request with only 628 bytes of header. The typical example illustrated at the Fig. 5. If we send one answer we need 658 bytes, but when we combine two answers in one request, only 688 bytes are needed, whereas the classic web application requires 1316 bytes to send two answers to the server. Thus, using Ajax the bandwidth is saved as shown on Fig. 5. The OT used at the figure represents old technology or classic technology. NT is used to represent new technology or Ajax technology.

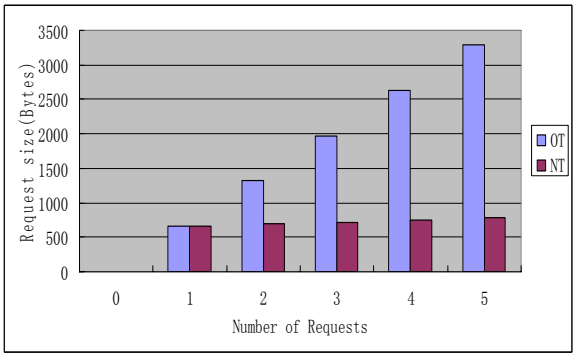


Figure 5. Request size comparison between classic web application technology and Ajax technology

V. CONCLUSION

In applications that have a significant part of each page containing content that is identical in multiple page requests, using Ajax-style methods to update only the relevant parts of a web page can bring significant bandwidth savings and reduce loading time. Caching

method also allows efficient usage of the bandwidth using aggregation of requests and reducing time of request using question data caching. That helps the student to do exam without any problem on bandwidth.

ACKNOWLEDGMENT

This work was supported in part by a grant from Hunan Province Natural Science Fund (No. 07JJ6122), Postdoctoral Science Fund of Central South University.

REFERENCES

[1] Yen-Ting Lin, Yi-Chiun Chi, Lien-Chien Chang, Shu-Chen Cheng and Yueh-Min Huang "A web 2.0 Synchronous Learning Environment using Ajax," Proceedings of the Ninth IEEE International Symposium on Multimedia Workshops, ISBN:0-7695-3084-2, Pages 453-458, 2007

[2] http://www.interaktonline.com/Support/Articles/Details/AJAX:+Asynchronously+Moving+Forward-How+does +AJAX+work%3F.html?id_art=36&id_asc=308, 13 October 2009

[3] Jesse James Garrett, "A New Approach to Web Applications," January 2005 http://www.adaptivepath. com/publications/essays/archives/000385.php, 12 0ctober 2009

[4] Matthew J. Travi, "Response Web Application Using Ajax," unpublished.

[5] Ridwan Sanjaya and Chaiyong Brahmawong, "Distance examination using Ajax to reduce web server load and student data transfer," Forth International Conference on eLearning for Knowledge-Based Society, 24thSouth East Asia Regional Computer Conference, Sunday, November 18, 2007.

[6] http://www.interaktonline.com/Support/Articles/Details/AJAX:+Asynchronously+Moving+Forward-How+does+AJAX+work%3F.html?id_art=36&id_asc=308, 13 October 2009

[7] http://articles.sitepoint.com/article/build-your-own-ajax-web-apps, 19 October 2009

[8] L.D. Paulson, "Building rich web applications with Ajax," Computer, IEEE, vol.38, no.10, Oct.2005, pp.14-17

[9] Tina Schmidt, "Routing and packet forwarding," September2008, http://www14.informatik.tu-muenchen. de/konferenzen/Ferienakademie08/talks/tina_schmidt/paper_schmidt_tina_routing_and_packet_forwarding.pdf, 20 October 2009

[10] http://web.uettaxila.edu.pk/CMS/seSPbsSp09/notes%5CTCP,IP,UDP%20Headers%20&%20TCP%20State%20Transition%20Diagram.pdf(TCP-IP-UDP), September, 2009

[11] http://raider.muc.edu/~kirchmjf/SP2004/cs360/EthernetFrame.htm, 15 October 2009

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 175-178

E-Commerce Trust Model Based on Perceived Risk

Ruizhong Du [1], Xiaoxue Ma [2], and Zixian Wang [3]
[1]Institute of Network Technology;Hebei University; Baoding, China ;
durz@mail.hbu.cn
[2]Computing Center;Hebei University; Baoding China ;
[3]Modern education technical center;The Central Institute For Correctional Police; Baoding China;
bd_wzx@126.com

***Abstract*—in the popularity of the Internet, e-commerce more and more into people's daily lives. E-commerce is different from the traditional forms of face to face trading patterns, which requires the users to have a good faith transaction. To detect the user's trustworthiness, models of reputation system have been proposed to the user's credibility to determine whether a transaction is feasible. Based on the previous reputation models and research, e-commerce trust model based on perceived risk(ECTMPR) is proposed in this paper. In the model perceived risk will be divided into six dimensions, by calculating the user's direct trust degree and recommendation trust degree, integrated into an end-user trust degree. By comparing the level of perceived risk and confidence levels to determine whether a transaction is carried out. Simulation experiments about malicious users and fraudulent users, show that the model can effectively improve the success percentage of e-commerce transactions, thus reducing losses.**

***Index Terms*— e-commerce, perceived risk, reputation, subjective trust**

I. INTRODUCTION

With the development of information technology, more and more people use the Internet to conduct transactions which changes the traditional trading patterns. But the Internet can't guarantee the credibility of the interaction parties, if the choice is not proper, the interaction could be failed or losses some economic, so the user must bear a certain degree of transaction risk. In order to solve such problems, the concept of reputation system has been proposed[1-3] and the model of reputation system is given.

As an effective evaluation mechanism of user behavior, reputation system has become a research hotspot in recent years. Average Reputation Model[4] use the cumulative averaged value as the user's reputation of his historical evaluation, so it does not take into account the affect of different periods. Beta Reputation Model[5] divides the historical evaluation into positive and negative, according the evaluation of these historical acts to calculating the expected reputation value of the next transaction. Such as eBay[6] who has the above calculation method about the reputation of the transaction value.

Perceived risk originally was introduced to consumer behavior from psychological[7] by Bauer (1960) who was from Harvard University. In the existing trust models, there is little model has considered the impact on trust evaluation of user's subjective perception factors. Therefore, the Perceived risk in e-commerce is introduced in trust evaluation, decision-making in the transaction with a trust degree of an entity to guide trading decisions, dealing with the perceived risk is assessed to provide an objective reference for decision-making, and to improve e-commerce transactions between the entities and the entities success rate.

II. ECTMPR MODEL BUILDING

Def1: Entity: In e-commerce system, the entity is the individual which request or provide service.

Def2: Risk of Perceived Degree: The expected quantification of uncertainty and adverse consequences probability which subjective perceives by entity, it is expressed by PR.

Def3: Direct Trust Degree: The confidence value which is obtained through a direct interaction of entity i and entity j, it is expressed by $DT_{i,j}$.

Def4: Recommended Trust Degree: The confidence value which is obtained through comprehensive calculation of entity i and multiple Direct Trust Degree about entity j in the system, it is expressed by $RT_{i,j}$.

Suppose there are two entities: entity i and entity j in e-commerce systems, the transaction process as follows:

Step 1: transaction request entity i calculates this transaction's perceived risk degree, in accordance with the value magnitude; and then sends transaction request information to the other entities of the system;

Step2: the entity which can provide service after received this transaction request sends a service response message to the entity i;

Step3: the entity i choose one provider from response entities, it is expressed by j, and then sends request information to its own trusted entity for the credibility value;

Step4: after received the request, service referral entity will feedback the direct trust information about entity j to the entity i;

Step5: entity i will receive direct trust degree on entity j, then uses it to get overall trust degree about the node j;

Step6: entity i uses overall trust degree and the perceived risk degree to decide whether transact with entity j which was chose on step 4. If the transaction is approvable, then do step 8; If it not agree to trade, then do

step 4 and repeats do steps 5 ~ 7 to evaluate the trust degree of transaction request entity.

Step 7: when the transaction end, based on the trading results update the direct trust degree of request node to service node.

A *Perceived Risk*

Perceived risk is psychological feelings and subjective understanding of consumers which are about various objective risks in the trading. Kotler [8] said: that consumer change，postpone or cancel the deal is affected on to a largely extent.

In this model, we pull the concept of perceived risk degree in the evaluation process of credibility; integrate multiple risk dimensions, to calculate the perceived risk degree. In e-commerce transactions, the buyer's subjective judgment, commodities' characteristics, as well as the seller' acts are factors that can influence the perceived risk.

2.1.1 Perceived risk dimensions divided

The definition of perceived risk by Bauer includes two factors: uncertainty and adverse consequences. The adverse consequence is the size of the loss by the consumer's subjective perceives when the purchase is negative. According to the different manifestations of adverse consequences, there are multiple dimensions of perceived risk criteria for the classification. Anne-Sophie Cases [9] use eight dimensions to measure the online perceived risk: financial risk, payment risk, performance risk, delivery risk, time risk, social risk, sources risks and privacy risks. In the empirical study of Chinese consumer online shopping Integrated multi-factor, we define six dimensions in the model: economic risk, time risk, functional risk, service risk, social risk and privacy risk.

2.1.2 Quantify the perceived risk

According to this model, entities can choose various dimensions of perceived risk according to their own subjective determination, this model correspond different trends to select different parameters δ_i.

Following formula (1) to calculate the perceived risk degree for each trading PR:

$$\mathrm{PR} = \sum_{i=1}^{n} \delta_i \otimes \mathrm{PL}_i \otimes \mathrm{IL}_i \qquad (1)$$

n: Perceived risk dimensions; PL_i: The possibility of occurring consequence i after the transaction.

IL_i: The severity of occurring i; α_i: The parameters of perceptions trends after i occurred.

PL_i Can be provided by the nodes or recommended reference information.

IL_i can be distributed according to the extent of adverse impact caused by adverse consequences

B *Calculation of confidence.*

To calculate the value of direct trust, after the entity i transacts with entity j, evaluates the satisfactory degree which remark with variable $V_{i,j}^k$, that the interval is defined at [0, 1]. After obtaining the score that entity i for entity j, then combines with historical direct trust value $DT_{i,j}^{k-1}$, we can get the direct trust value of entity i to entity j after k-transaction.

$$DT_{i,j}^k = \begin{cases} v_{i,j}^k, & k = 1 \\ \partial DT_{i,j}^{k-1} + (1-\partial) v_{i,j}^k * c_{i,j}^k, & k \geq 2 \end{cases} \qquad (2)$$

Among them, parameter ∂（$0 \leqslant \partial \leqslant 1$） is historical factor, that when calculating the direct trust value, the weight of the historical credibility. Parameter $c_{i,j}^k$ is said transaction context for k-transaction, in this paper express that the value of this transaction.

In the above formula, the sustainability of the entity's conduct doesn't consider incompletely. So, based on above formula, we introduce λ that is increasing with the trading number. Suppose $\lambda = \sqrt{k(k+1)}$, amend the above formula:

$$DT_{i,j}^k = \begin{cases} \lambda v_{i,j}^k, & k = 1 \\ \lambda(\partial DT_{i,j}^{k-1} + (1-\partial) v_{i,j}^k * c_{i,j}^k), & k \geq 2 \end{cases} \qquad (3)$$

So, if entity wants to get a higher value of credibility, they must have been praised continuously.

To calculate the recommended trust degree, we use entity i to collect directly trust degree about j within the system, then calculate to get recommend trust degree on j:

$$RT_{i,j} = \sum_{k=1}^{N} R_{i,k} * DT_{k,j} / N \qquad (4)$$

N: the number of recommend entity; $R_{k,j}$: the direct trust degree of entity k to entity j;

$R_{i,k}$: Here we use it as entity k's recommend trust coefficient; this article assumes the credibility of the initial recommend entity is 0.5.

Integrated direct trust degree and recommendation trust degree, we get the overall trust degree formula:

$$T_{i,j} = \beta DT_{i,j} + (1-\beta) RT_{i,j}, \quad (0 \leqslant \beta \leqslant 1) \qquad (5)$$

β: The weight value of direct trust degree in overall trust degree. Combine perception risk PR with the trust lever $\mathrm{TL}_{i,j}$ which is entities I to entity j , determining whether transact with entity j.

III. Simulation Experiment

In order to verify the validity of the model, based on query cycle simulator who use file-sharing mechanism this article achieve validity of the simulation model. To express the superiority of the model, this article compared the successful transaction percentage in the same simulation environment for the Average, Bate, and

eBay.

A Simulation experiments about Containment malicious conduct

This scenario is to verify the effectiveness that the models can hold back malicious. Table 1 shows relevant parameters and their values.

Table 1 relevant parameter and their values

Description	Default value
β	0.6
∂	0.2
the number of users in the system	200
the proportion of malicious users	[0,0.5]
simulation cycles	100
the probability of transaction	80%
request of transaction mode	Random

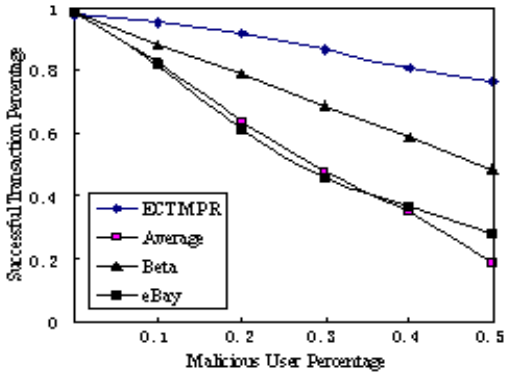


Fig 1 the rate of malicious users with successful transaction

Fig1 shows the successful transaction percentages change with the proportion of malicious nodes, when the malicious user attacking. When the system does not have a malicious node, the successful transaction percentage is 100% for the four kinds of models. With the increasing of malicious nodes, the four models' successful transaction percentages are turning down. But the perceived risk can feel the level of risk and then reduce the probability of transaction, which can effectively prevent cheat ant improve the success transaction percentage of an honest user.

B Simulation experiments about Containment Deceptive user

This scenario is to verify the effectiveness that a model can hold back deceptive user. Table 2 shows relevant parameters and their values.

Fig2 shows change with the proportion of malicious nodes, when the deceptive user attacking. With the increasing of deceptive users, the honest users will meet more and more deceptive users, so the successful transaction percentages are declining. However, ECTMPR model can provide a better rate of successful transaction.

Table 2 Relevant parameters and their values

Description	Default value
β	0.6
∂	0.2
the number of users of the system	200
the proportion of deceptive users	[0,0.5]
the threshold provide honest behavior	0.5
the threshold provide deceptive behavior	0.7
simulation cycles	100
the probability of transaction	80%
request of transaction mode	Random

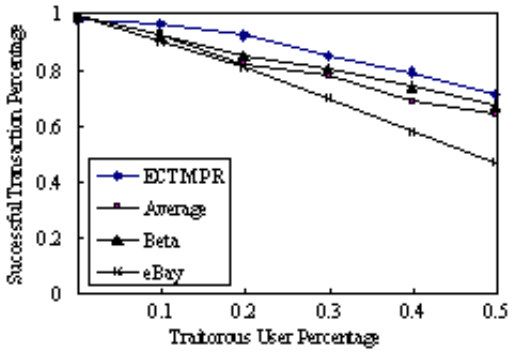


Fig 2 the rate of deceptive users with successful transaction

IV. CONCLUSIONS

There is a little model has considered the users' subjective perception factors who have impact in. This article introduces user-perceived risk to trust evaluate and provide an e-commerce trust model. By combining direct trust degree, recommendation trust degree and other influence factors we get comprehensive trust degree of entity. Then with the user's perceived risk degree to guide trading decisions. Through the simulation experiments, and compare with some models, the model can more effectively improve the successful transaction percentage for honest users.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (60873203), Natural Science Foundation of Hebei Province (F2008000646), Educational Commission of Hebei Province （ZH2006006）,Natural Science Foundation of Hebei University (2008Q51)

REFERENCE

[1]. Resnic P, Kuwabara k, et al. Reputation systems. Communications of the ACM 2000,43(12):45-48

[2]. Chang E, Dillon T S, Hussain F K. Trust and Reputation Relationships in Service-oriented environments. 3rd International Conference on Information Technology and Applications. Sydney:IEEE,2005.4-14

[3]. LIANG Z Q, SHI W S. PET: A Person-alized trust model with reputation and risk evaluation for P2P resource shari-ng",the 38th Hawaii International Conference on System Science,2005

[4]. Jurca R, Faltings B. Towards Incentive-Compatible Reputation Management. Lecture Notes in Computer Science. Volume 2631/2003,13-24

[5]. Jφsang A, Ismail,R.The Beta Reputation System. 15th Bled Electronic Commerce Conference,2002.324-337

[6]. Resnick P, Zeckhauser R, Swanson J, Lockwood K. the Value of Reputation on eBya:A Controlled experiment. Experimental Economics. Volume 9, Issue 2, 2006, Page 79-101

[7]. Bauer R.A. Consumer Behavior as Risk Taking. Dynamic Marketing for a Changing World. 1960:389-398

[8]. Kotler P. Marketing management : Analysis , planning , implementation, and control [M] . 9th ed. Canberra: Prentice Hall , 1999

[9]. Anne-Sophie Cases. Perceived Risk and Risk Reduction Strategies in Internet Shopping. The International Review of Retail, Distribution and Consumer Research. 2002(10)

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 179-182

Implementation of the Authorization Management with RBAC in the Usage Control Model

Hui Cai[1], and Peiwu Li[2]
[1] Nanchang Hangkong University,School of Software, Nanchang, P.R.China
jxzsch@163.com
[2] Scientific Research Office of NIT, Nanchang, P.R.China

***Abstract*—The usage control which involves traditional access control, trust management and digital right management is a comprehensive model. For the reasons included high abstract and the authorization hard to be managed, a new model was putted forward to administrate usage control mode1——RUCON model. Six functions of authorizations, which are user-role auto-assignment based on role-rule, role-hierarchy assignment, permissions -role assignment, usage-rule, obligations and conditions were defined in the new model. In order to simplify the management the new model adopted the advantages of the RBAC model. In the RUCON model the assignment and revocation of right are administrated automatically through the medium of roles. The formal description of the RUCON model was given as well.**

***Index Terms*—access control; usage control; authorization; ABC model; assignment and revocation of the right**

I. INTRODUCTION

As a unified framework for next generation access control, the usage control model (UCON) [1] has three factors of decision which are Authorization, oBligation and Condition. Continuity of decision and mutability of attributes are defined in the model as well. There are many effective results about conceptual model of usage control. The ABC model [2] proposed by professor Sandhu is a good example. The ABC core model for UCON lays the foundation for next generation access controls that are required for real world information and systems security, but there are some shortcomings about the model. For example, research about the management of authorization is very poor now. It necessary to be further discussed in the future.

Role-based access control (RBAC) has a good management model. The course of authorization became more flexible after the factor of role add in. it also can adapt to a variety of access control through the agency of the role of inheritance. RBAC system has been studied and discussed recently. The assignment and revocation of the right can be achieved in the closed or the network environment. Even some problems in continuity also have good solvable strategies [3]. In this paper, a management model which based on the conceptual model of UCON and the idea of RBAC is described to achieve the auto-assignment of rights. The model defines a mechanism to determine seniority among rights. When certain conditions hold, the model also allows dynamic revocation of assigned right. The paper also shows how to use the model to express the policy of authorization.

The paper is organized as follows. In section 2 we summarize related research about UCON model. Section 3 describes our model. In sections 4 shows how our model can be used in one real life examples. In section 5 we conclude the paper and touch on issues that we have not explored in this paper, though they are closely related to the topic discussed.

II. THE USAGE CONTROL MODEL

The ABC models consist of eight core components. They are subjects, subject attributes, objects, object attributes, rights, authorizations, obligations, and conditions. Authorizations, obligations and conditions are functional predicates that have to be evaluated for usage decision. They enrich the access control policy and improve the shortcomings and deficiencies of traditional access control when they work in an open network environment. Two features of the ABC model which are continuity of decision and mutability of attributes are the essential difference between traditional access control and usage control.

III. THE RUCON MODEL

As shown in Figure 1, the RUCON model is a management model which can automatically achieve the assignment of the rights.

RUCON model consists of 11 elements, which are subject, object, subject attributes, object attributes, Usage-Rule, oBligations, conditions, permissions, session,

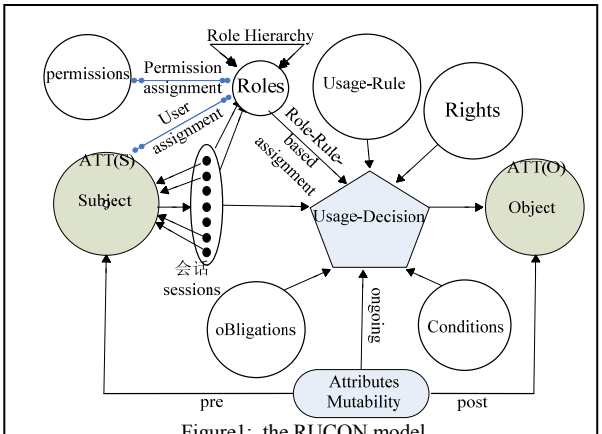


Figure1: the RUCON model

role and the Role-Rule. The idea that uses the policy of role-based access control to achieve the management of usage control has been proposed in some literature. In this paper the authorization in UCON model will be further discussed. To a certain extent RUCON model can realize the continual assignment and revocation of the right and hit the purpose that simplify and administrate the course of authorization.

A. Components of the model

Definition 1: Role-Rule is a set of rules based which the system completes the auto-assignment during the process of dynamic authorization. A user will be automatically assigned the appropriate role when his attributes meet with the Role-Rule which has an attribute expression [4]. The rule is defined by the people who have the permission in the enterprise. To define this set of rules you must consider the user's own attributes and a variety of constraints. Users' attributes are provided along with the authentication information or can be fetched from databases. The Role-Rule based auto-assignment is used in the process of continual authorization.

Definition 2: Usage-Rule is a constraint used to determine the subject can access the object or not after the subject obtain a role and a Boolean type will be returned. The authorization will be automatically revoked by the Usage-Rule if a FALSE returned.

Definition 3: The 6 authorization functions in RUCON model are: Role-Rule-based assignment, inheritance-assignment, permissions assignment, Usage-Rule, obligation and condition. The first three functions are used to realize assignment of the right and the last three functions are used to realize revocation of the right.

B. Authorization policy In the model

In the RUCON model, according to the security policy the administrator defines a set of Role-Rule based on the user's attributes for the system. The system achieves the user-role assignment automatically by parsing the Role-Rule and monitoring the changes of user's attributes. In this way the problem about role identification in the open network environment is solved. And then the system completes the authorization dynamically considering constraints such as Usage-Rule 、 obligation and condition.

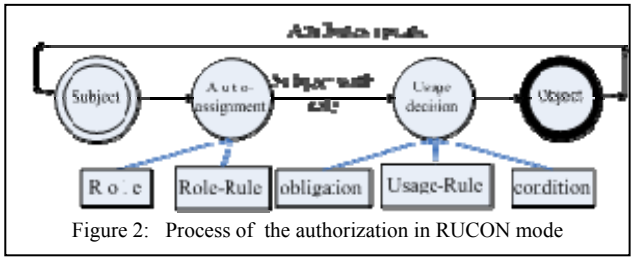


Figure 2: Process of the authorization in RUCON mode

As shown in Figure 2, the process of authorization is continued and cyclical. The change of user's attributes is very important because the attributes directly affect the decision of authorization and the side effects after authorization probably change the attribute values. [5] studied in detail about the attribute mutability in usage control. The attribute has some variations which are mutability for exclusive/inclusive attributes、mutability for consumable/creditable attributes 、 mutability for immediate revocation 、 mutability for obligation and mutability for dynamic confinement. These variations can describe almost the changes of user's attributes and the attributes are expressed by attribute expressions. According to the security policy defined before, if the seniority level of user's Attribute_Expression is equal or senior to the Role-Rule's Attribute_Expression the user will be automatically assigned a role corresponded to Role-Rule:

Attribute_Expression (ATT(U)) ≥ Attribute_Expression (Role-Rule) ⇒ Role.

The authorization in the model can be expressed as:

allowed(s,o,r) ⊆ S× O× P× Ri× Usage-Rule× B× C.

The 6 functions defined in the model are used to administrate the authorization, in which permissions assignment 、 obligation and condition are same as ARBAC97 model and ABC model. Role-Rule-based assignment、inheritance-assignment and Usage-Rule are describe as follow:

1. Role-Rule-based assignment

[1] user-role auto-assignment: can-assign (Role-Rule, Roles)

If the seniority level of user's Attribute_Expression is equal or senior to the Role-Rule's Attribute_Expression the user will be automatically assigned the role corresponded to Role-Rule.

[2] user-role revocation: can-revoke (¬ Role-Rule, Roles)

When the seniority level of user's Attribute_Expression is junior to the Role-Rule's Attribute_Expression the system will automatically achieve the revocation.

2. Inheritance-assignment

A user can be assigned the role corresponded to Role-Rule if he has the proper attribute expression which is senior than Role-Rule's. Two attribute expressions can be described as comparable only if :1) they have the same construction.2) they satisfy the same constraints.

An attribute expression may be hierarchically related to one or more attribute expressions only if they can be describe as comparable. We use " ≥ " to express the seniority level between attribute expressions.

Definition 4: The Role-Rule1 with Attribute_Expression1 is senior to the Role-Rule2 with Attribute_Expression2 only if Attribute_Expression1 is senior than Attribute_Expression2.

A senior rule inherits all the roles produced by any of its junior rules and roles produced by senior rules have high seniority level. Because of these seniority levels a hierarchy of role is created. We call it induced role hierarchies (IRH), which indicates a hierarchical relationship between roles and these roles depends on the

Role-Rule. Some times, the security administrator may encounter some other hierarchies such as the given role hierarchies (GRH). The GRH is generated by the security policy. The permissions are inherited in the GRH from bottom to top and the GRH includes all the roles belong to IRH in the ideal case: $\text{Roles} \in \text{GRH} \supseteq \text{Roles} \in \text{IRH}$.

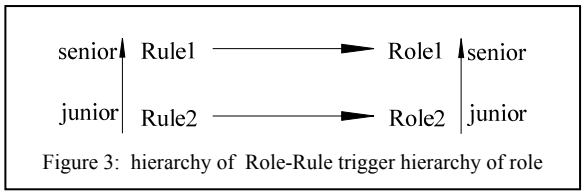


Figure 3: hierarchy of Role-Rule trigger hierarchy of role

As shown in figure 3, role1 is produced by rule1 and role2 is produced by rule2. Rule1 is senior to rule2, so the role1 is senior to role2 and the user correspond to rule1 can inherit role2. The IRH is induced.

IRH hierarchy corresponds to the user-role assignment and GRH hierarchy corresponds to permission-role assignment. in the IRH a junior role inherits all the users who had been assigned to its ancestor(for the user-role assignment) and in the GRH a senior role inherits all the permissions that had been assigned to its descendants (for permission - role assignment), Figure 4 shows the relationship between these two kinds of inheritance:

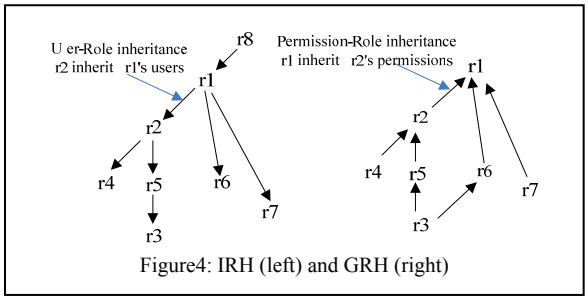


Figure4: IRH (left) and GRH (right)

Under the ideal circumstances, the roles in IRH and GRH are corresponding with each other:

$(\forall \text{ Role in GRH}) \in \text{IRH} \cap (\text{Role in IRH}) \in \text{GRH}$

As shown in figure 3 and figure 4 we can describe the inheritance-assignment like this:

a. user-role assignment: user1 can also be assigned role2 by the IRH (role2 inherits role1's user1).

b. permission-role assignment: role1 inherits role2's permission by the GRH because role1 is senior to role2.

3. Usage-Rule

According to the process in figure 2, after the subject obtained its role the system check the role of subject and object. If Seniority Level of (subject)<Seniority Level of (object) or the role of subject not meets the Usage-Rule then the revocation would be triggered: $\text{Stop}(s,o,r) \Leftarrow \text{Role}(s) < \text{Role}(o)$.

4. Constraint in UCON

In RBAC model the major constraints are static separation of duty and dynamic separation of duty. The first occurs at the moment that assigns permission to user by administrator:

$$\bigcap_{Z \in SSD} \text{can-assign}(x,y,z) = \Phi .$$

The dynamic separation of duty occurs during Role-Rule-based assignment. In a session, the system checks the role assigning to subject as long as the attribute changed. If the role conflict with the previous one the authorization is stop:

$$\bigcap_{Z \in SSD} \text{can-assign}(x,y,z) \Phi .$$

$$\bigcap_{Roles \in DSD} \text{can-assign}(\text{Role-Rule}, \text{Roles}) = \Phi .$$

5. Continuity of decision and Mutability of attribute

As shown in figure 2, the process of authorization is continued and cyclical. The attribute is also constantly changing throughout the process. In a session the attribute of subject may be changed under the side effect after authorization and then the attribute of subject update: Update(ATT(s)). The change of attribute results in that the role obtained by subject will be changed during next process of Role-Rule-based assignment. After the role changed a decision would be made by system continuously. Dynamic separation of duty is a good example.

IV. CASE

An online entertainment store provides movies, games documentary films and music, etc. The permission provided is rated according to a hypothetical rating system as shown in Table 1.Using RBAC terminology, levels correspond to roles, which, in turn correspond to permissions. Roles are totally ordered in this example. When users logon, the attributes they provide determine the highest level they can obtain.

TABLE I. ROLE-PERMISSION TABLE

Rating level	Permission	Corresponding Role
1	only can read the introduction of movies, games and music	viewer
2	in addition to the permissions of a view the ordinary member can also enjoy the movies, music, etc online	ordinary member
3	in addition to the permissions of a ordinary member the VIP member can download the resources wanted, such as games and movies.	VIP member

.For the sake of the discussion, we will consider 2 attributes: the account-balance and consumer-credit of the users. If his\her account-balance is equal to zero then the user is a viewer. A ordinary member will upgrade to VIP member if his\her consumer-credit is equal to or greater than 100. We defined as follows:

a) Attribute::=account-balance | consumer-credit

b) Attribute_Value::= numerical value N

c) Role::= viewer | ordinary member | VIP member

Since no constraints or conditions were specified, the following rules are produced:

Rule1:: (account-balance=0) $\Rightarrow$ viewer.

Rule2::(account-balance>0) $\cap$ (consumer-credit $\geq$ 0) $\Rightarrow$

ordinary member.

Rule3: (account-balance>0) $\cap$ (consumer-credit $\geq$ 100) $\Rightarrow$ VIP member.

1. Role-Rule-based assignment

can-assign(rule1,viewer) $\Leftarrow$ account-balance=0

can-assign(rule2,ordinary member) $\Leftarrow$ (account-balance>0) $\wedge$ (consumer-credit $\geq$ 0)

can-assign(rule3,VIP member) $\Leftarrow$ (account-balance>0) $\wedge$ (consumer-credit $\geq$ 100)

can-revoke($\neg$ rule1,ordinary member) $\Leftarrow$ account-balance=0

can-revoke($\neg$ rule2,VIP member) $\Leftarrow$ consumer-credit<100

2. Inheritance-assignment

Rule1 corresponds to Attribute_Expression1 :

.account-balance=0

Rule2 corresponds to Attribute_Expression2 :

(account-balance>0) $\cap$ (consumer-credit $\geq$ 0).

Rule3 corresponds to Attribute_Expression3 :

(account-balance>0) $\cap$ (consumer-credit $\geq$ 100).

Obviously, the follow expression is correct:

Attribute_Expression1< Attribute_Expression2<

Attribute_Expression3

According to definition 4 we can get the relationship between the rules: rule1<rule2<rule3. Through the inheritance-assignment the senior rule can get the role produced by junior rule. For example, a user upgrade to VIP member because his\her attribute expression satisfies the rule3. At the same time, since rule3 is senior to rule2 the user can also be assigned to the role ordinary member, which correspond to the rule2.

3. Usage-Rule

When an ordinary member want to download a documentary films he\she will find he\she has no the permission. The process of authorization is described as follow:

The object (documentary films) need role VIP member and the subject's (the user's) role is ordinary member which junior to the required. According above definitions the below will be implemented automatically: Stop(s,o,r) $\Leftarrow$ Role(s) <Role(o).

4. Continuity of decision and Mutability of attribute

In the case, the account-balance is a consumable attribute and the consumer-credit is a creditable attribute [5]. We give some Assumptions:

After the user logon his\her account-balance reduce 1 per one hour.

After the user logon his\her consumer-credit increase 10 per one hour.

User1 has 2 points in his\her account-balance and 90 points in his\her consumer-credit. After one hour from user1 logon his\her consumer-credit reach to 100 points and his\her attribute update automatically:

Update(ATT(User1)) $\Rightarrow$ consumer-credit=100.

Under the effect of Role-Rule user1 is assigned the role VIP member and then he\she can download the resources online. After two hours from user1 logon his\her account-balance reduce to 0 point and the attribute update automatically:

Update(ATT(User1)) $\Rightarrow$ account-balance=0.

Under the effect of Role-Rule user1 is revoked the role VIP member and assigned the role viewer. Then user1 only can read the introduction but has no permission to enjoy movies online or download the resources.

V. CONCLUSION

We have described a model to implement the authorization in the usage control model. The factor of role increases the authorization steps and additional resources of system are consumed, but the management of authorization in usage control model becomes more simple and flexible. We believe that our model will be useful in automatic authorization.

REFERENCES

[1] JaehongPark and Ravi Sandhu,Usage Control: A Vision for Next Generation Access Control[C], MMM-ACNS,2003.

[2] Jaehong Park and Ravi Sandhu,The $UCON_{ABC}$ Usage Control Model[J],ACM Transactions on Information and System Security, Vol. 7, No. 1, February 2004, Pages 128~174.

[3] Axel Kern, Claudia Walhorn. Rule support for role-based access control[J], Proceedings of the tenth ACM symposium on Access control models and technologies, June 01-03, 2005

[4] Mohammad A. Al-Kahtani, Ravi Sandhu, A Model for Attribute-Based User-Role Assignment[C], Proceedings of the 18th Annual Computer Security Applications Conference, p.353, December 09-13, 2002

[5] ParkJ, ZhangX, SandhuR. Attribute mutability in usage control [J] Annual IFIP WG Working Conference on Data and Applications Security, 2004, 11(3):1~12.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 183-186

Research on E-Learning System and Its Supporting Products:A Review Base on Knowledge Management

Zaiwen Wang[1] , Yang Zhao [2] , and Yanping Liu[2]
[1]School of Economics ,Beijing Technology and Business University,Beijing, P.R.China
sxcdwzw@163.com
[2]School of Economics & Management,Beijing Jiao tong University,Beijing, P.R.China
zhaoyang456@sohu.com

***Abstract*-Knowledge management(KM) is supported by many strategies such as business intelligence, collaboration, document management and e-learning (E-L). E-L has played more and more important role of all the technologies in the supporting KM. The review of basic technologies that support E-L will be in favor of further study on E-L. Base on KM, paper makes a review about the relationship between E-L and KM and products that support the design and operation of E-L system. Especially, we introduce the current development condition of E-L in China. With these reviews of E-L, we find that more theory about corresponding products is needed to guide the design, delivery, and implementation of E-L.**

***Index Terms*-E-Learning; system; supporting products; Knowledge Management**

I. INTRODUCTION

In the new global competitive landscap, firms succeed not because they have superior control over scarce resources, but because they are able to learn and touse this learning more effciently than others. In this situation, KM system applying new advanced technologies will meet the need of organizations to enhance their learning ability. A successful KM system depends on many key factors, such as strategies, system structure, supporting technologies and operation tools. In interviewing a number of distance learning subject matter experts, scholars like Welsh found that four themes will characterize the landscape of E-L during the next several years[1].

The current state of the art of E-L in organizations reveals that many of these predictions are already coming to fruition. So this paper tries to show the differences and the similarities between what has been done in the east and in the west through a literature review of the development of technologies base on KM.

II. LITERATURE REVIEW OF E-LEARNING

A. Relationship between KM and E-L

To study and review products supporting to E-L system, we must first find out the guiding theory of E-L system. With the guiding theory we can get the basic direction for our further study. Many scholars have found that the emphasis on E-L has become shifting to "performance support" with the integration of KM capabilities[2].

Many corporations are discovering that E-L has many of the same attributes as basic KM processes and thus can be as a tool for KM[3]. Based on Neumann research of E-L and cooperation as elements of KM, E-L makes an important contribution to accessibility, transparency and maintenance knowledge[4].

We can conclude that KM is the premise and operational platform of E-L system and the E-L is the key technology and tool supporting KM.

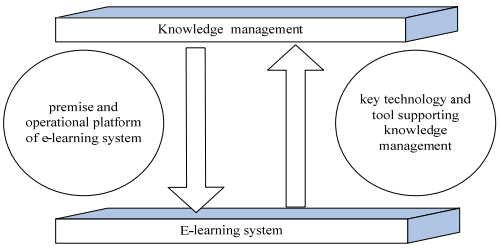


Figure 1 Relationships between KM and E-L

Figure1 shows the relationship between KM and E-L.

B. Critical Success Factors for E-Learning System

Why we make this review is that technology has become the most important factor among critical success factors (CSFs) that influence the successful operation of E-L system. In this part we will introduce the CSFs put forward by scholars from both western countries and China.

From the view of the research, we can infer that there are various factors that can influence the success of an E-L system.Table1 shows the CSFs proposed by the three western scholars. And Table 2 shows the main opinions of Chinese scholars on the CSFs of E-L.

From table1 and table2 we can find that technology is widely perceived as a CSFs that influence the successful operation of E-L system. So it is necessary to pay more attention to the development of technology that support E-L system.

AP-PROC-CS-09CN005

TABLE Ⅰ

CSFS OF E-L PROPOSED BY WESTERN SCHOLARS[5][6][7]

Factor \ Scholar		Gascoet al	Selim	Henry
organization	flexibility in management	√		
	control mechanisms	√		
	communication	√		
	instructor's attitude		√	
individual	active participation	√	√	
	learners motivation		√	
technology	use of advanced technologies	√	√	√
	internet environment		√	
culture	culture establishment		√	
	establish mutual trust			
content				√
service				√

TABLE Ⅱ.

CSFS OF E-L PROPOSED BY CHINESE SCHOLARS[8][9][10][11]

Scholar / Factor	Liang et al.	Zheng et al.	He et al.	Li Y. et al.
flexibility of management				
organization			√	
instructor's attitude	√	√		√
participation of key personnel				
learners motivation	√	√		√
technology	√	√	√	√
content	√	√	√	
available resource				√
Learning strategy		√		
internet environment				√
culture				

Different scholars of western counties have put forward different technologies to design E-L system. There technologies enhance the learning efficiency and develop the KM of organization.

Virtual Reality Modeling Language (VRML) is the 3D language of the Web. Its purpose is to provide information to web pages in a three dimensional format. Problems facing the widespread adoption of VRML in E-L are the necessity for a client-side plug-in to be installed on the learner's computer. This problem hopes to be solved by the development of international standards for VRML, currently being drafted by the Web3D Consortium[9].

Portal is a learning technology put forward by Brandon Hall[10]. He offers an explanation of learning portals that "Learning Portals are web sites that provide a combination of courses, collaboration and community. Initially set up with ecommerce for the individual purchaser with a credit card, most portals have plans to offer credits of some type for multiple registrations from a single organization.

Knowledge Tree is presented as a framework for adaptive E-L based on distributed re-usable learning activities that we are currently developing. The goal of KnowledgeTree is to bridge the gap between the information power of modern educational material repositories and the just-in-time delivery and personalization power of ITS and AH technologies[11].

Wikis, fully editable websites, are easily accessible, require no software and allow its contributors, in this case students, to feel a sense of responsibility and ownership. Wikis are everywhere, but, unfortunately, the online literature has not yet begun to focus enough on wikis[12].

From the analysis above we can find that the technologies of E-L changed and developed rapidly in western counties.

Ⅲ. REVIEW OF PRODUCTS THAT SUPPORTING E-LEARNING SYSTEM

Information technology is the most important supporting factors for the successful operation of e-learning system. By understanding the supporting technologies of e-learning, organizations can find out the basic operation rules of e-learning and can construct e-learning system with realizable technology. But the successful operation of an e-learning system needs tools and products to transfer knowledge to members of organization. These products provide effective and efficient access to information resources. In this section we will review main products of e-learning system that supports knowledge management of organization. There are many products supporting e-learning. Table3 collects main examples of e-learning products proposed by scholars of western counties and table4 shows the main products developed by Chinese scholars.

TABLEIII

E-LEARNING PRODUCTS DEVELOPED BY WESTERN SCHOLARS [12][13][14]

Product	Description
Linux operating system	The product is object-oriented projected to be easily customized for each type of Linux system installation. This product was made using the facilities of IDE Delphi. Because Delphi is object-oriented, this application inherits some classes corresponding to application that simulates Windows XP installation.
e-Learning Suite	Mainly developed for training sales workers, the eLearning suite, consisting of eTraining and Siebel Distance Learning, provides automated content management, methods of measuring learning, and course content delivery. These two applications can be purchased and deployed together or as stand-alone products.
Human Capital Management Suite	The suite includes trademark KP, Performance, KP Learning, and KP Content. Learners can create customized blended online learning curricula. Products test and track learner progress and activities.
TrainNet	Consisting of five modules, this virtual and integrated on-line learning system works for a variety of delivery modalities. It integrates full-screen video with live interaction, using audio conferencing, synchronized Web content, application sharing, embedded email, and whiteboard and Q&A features.
Vuepoint Learning System 3.0	Four modules make up this e-learning and content management system: a Web-based evaluation, teaching, and research tool; a student testing and course tracking program; a template-based content creator; and an off-line viewer for asynchronous learning. Such an integrated system allows accompany to save multiple authoring and licensing fees and to conduct real-time course management.
LMS (Learning Management System)	Learning Management Systems (LMS) is the starting point (or critical component) of any e-learning or blended learning program. LMS offers its greatest value to the organization by providing a means to sequence content and create a manageable structure for instructors/administration staff.[17] The functions of a complete LMS include the following functions: system management, user management, testing system, course management, study module, study collaboration and resource management.

TABLEIV

E-LEARNING PRODUCTS DEVELOPED BY CHINESE SCHOLARS [15][16]

Product	Description
Learning-assistant system based on JSP on Web	This system is an information process system. Its most basic function is to experience input information and classify, process and store the input information, and then change the input information into the needed information according to the fixed method. Finally it will sort out the information to users by output system. During this course, the most central part is the information processing, so different data determines different treatment methods and algorithms needed.
Individualized Online learning system	This system proposes a solution scheme of obtaining learning resources with individual characteristics based on Web Services technology. Utilizing this scheme, the learner can obtain the most needed learning resources from plenty of organizations when learners search learning resources at each client that offers this service.
E-learning system model based on affective computing	E-learning system model based on affective computing can resolve the problem of communication deficiency between the computer and users effectively. This model is e-learning system that is based on the facial expression recognition and speech emotion recognition technologies. This system is used to determine the acceptance condition of study according to the students' emotion signal. This system can be divided into two parts: user's information introduction and systematic learning guidance. User's information introduction includes interface agent, emotion calculate server and user's account database. The part of systematic learning guidance includes individualized guidance agent, course materials database and user's materials database.
Web—based collaborating learning supporting platform	Web-based collaborating learning supporting platform takes constructivism learning theory and systematology as the guide and utilizes technologies such as ASP, XML, DOM, ADO to set up a kind of web-based learning supporting platform under the environment of INTERNET. This system can establish good academic environment and can fully improve the learner's innovation ability. This system adopts three layers structure of B/S mode: the front is one circle browsers, the middle is WEB server and the back is a database server. ASP, ADO interface is the middleware between WEB server and database server. When users send out the requirement for reading ASP file to Web server through the browser, web server will carries out ASP file and finally web server will return the result back to users.

Ⅳ. THE TECHNOLOGY DEVELOPMENT TRENDS OF E-L SYSTEM

From the literature review, we have sum up the characteristics of E-L and the basic technologies of E-L system. The review brings us a picture of the actuality of E-L theory and application. E-L develops rapidly in recent years. The question is, what should E-L be in the future? In this section, we will discuss the development trends of E-L technology.

At present, in most organizations, KM and E-L are two fields that have different positions. E-L is applied as the transfer tool of knowledge for organizational learning

and training and KM is applied as a strategic partner with executive decision makers. E-L users need a suitable KM that can help them to obtain the kind of content they need together with as correct and complete information as possible and effective management of knowledge are critical issues to the success of E-L systems [16].

Because the technologies of E-L depend on the internet and information technologies, organization members who are not good with computer operation and who with little IT knowledge may be hesitate to use these technologies. This problem will battle the successful operation of E-L system and reduce the knowledge learning level of organization. Collaborative E-L will solve this problem. Collaborative E-L includes man-machine and man-man interactions. Now there is little room for the technical development of man-machine interaction in terms of E-L platform. So a great many owners turn their focus on how to build up a more perfect learning environment for collaborative E-L.

Ⅴ. CONCLUSION

With the increased demand for building and maintaining ongoing capabilities, E-L has played more and more important role among all the technologies in the supporting process of KM. The development of E-L system is an important strategy in implementing KM policy.

From perspective of KM of organization, this paper makes a review about E-L system, especially on the relation between E-L and KM, CSFs that influence the development of E-L system and advanced technologies that support the design and operation of E-L system. Considering the importance of China in developing the technology of E-L, we especially emphasized the current development condition of E-L systems in China. With the development of information technology, we believe that more advanced technologies that support E-L will be created and that the integration of E-L systems and KM systems will be possible. E-L will have a very promising future.

REFERENCES

[1] Welsh, E. T.,Wanberg, C. R., Brown, E. G., & Simmering, M. J.. "E-L: Emerging uses, empirical results and future directions. International Journal of Training and Development, 2003,Vol. 7, pp.245-258.

[2] MindSpan. "The future of E-L: an expanding vision". Mercer Management Consulting Study (IBM MindSpan),. 2001

[3] Rosemary H. Wild. "A framework for E-L as a tool for KM".Industrial Management & Data Systems Vol. 7, 2002,pp.371-380.

[4] Neumann, H., & Schupp, W. "E-L and cooperation as elements of KM". Standard Eisen, 123(9), 2003,pp.81–84.

[5] Jose L. Gasco, Juan Llopis and M. Reyes Gonzalez. "The use of information technology in training human resources : An E-L case study, Journal of European Industrial Training". Vol. 28 No.5, 2004,pp.370-382.

[6] Hassan M. Selim. "Critical success factors for E-L acceptance": ConWrmatory factor models Computers & Educatio. 2005

[7] Paul Henry. "E-L technology, content and services". Education Training, Vol. 43.No. 4, 2005,pp.249-255.

[8] Liang Mingbo, Qu Li, Jin Chunhua, The outlook of e-learning in enterprises,Commercail Research, Vol.21, 2005,pp.367-71

[9] Li Yahping, Zhang Yujing, Study on implement of training system based on e-learning in enterprises, China Soft Science, No.4, 2003, pp.70-74

[10] He Feng, Wei Shunping, Han Guangyan, Research on the implement and application of e-learning in enterprises, Economics, Vol.12, 2004, pp.32-41

[11] Byron Marshall, Convergence of knowledge management and e-learning: the GetSmart experience, Proceedings of the 2003 Joint Conference on Digital Libraries. 2003. IEEE.

[12] George Siemens, Learning Management Systems: The wrong place to start learning, http://www.elearnspace.org/Articles/lms.htm, Vol.22, 2004

[13] Jing Luan, Andreea M. Serban. "Technologies, products, and models supporting KM".New Directions for Institutional Research, No. 113, Springp, 2002,pp.85-104.

[14] Jing Hong, Zhan Haisheng, Acquirement of individualized learning resources base on Web Services, Distance Education in China, No.8,2006,pp.63-65

[15] Ma Xirong, Liu Lin, Sang Jing, The e-Learning system model based on affective computing, Computer Science, Vol.32, No.8, 2005, pp.131-133

[16] Shao Jie, Design and implementation of a Web—based collaborating Education Supporting Platform, Modern Distance Education, No.4, 2005, pp.43-46

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 187-190

A Part-of-speeach Tag Sequence Text Zero-watermarking

Lu He [1,2], LingYu Zhang [1], GuangPing Ma [1], DingYi Fang [1], and XiaoLin Gui [2]
[1] School of Information Science & Technology, Northwest University, Xi'An, China
helu1977@yahoo.com.cn
[2] School of Electrical and Information Engineering, Xi'An JiaoTong University, Xi'An, China

***Abstract*—Zero-watermarking technology without modifying the carrier's data, resolved the imperceptibility of watermarking. Since natural language is lack of embedding space, zero-watermarking is suitable. Without modifying carrier text passive attack does not works, and only active attack may destroy the zero-watermarking. In this paper, we proposed a novel text zero-watermarking algorithm. Words, which are corresponding to one special POS tag sub-sequence pattern, were extracted as candidate for zero-watermarking. Again, words were chose under the control of chaotic function. Since there are so many possible permutation of POS tags sub-sequence and chaotic function has sensitive dependence on initial conditions or the key, the robustness was guaranteed. Experiments results show that it is almost impossible destroyed the zero-watermarking by stochastic synonym substitution and sentence transform and the results also show the robustness is better than other text zero-watermarking algorithms.**

***Index Terms*—Text Watermarking; Zero-watermarking; Natural Language Watermarking; Information Hiding; Chaos**

I. INTRODUCTION

With the rapid development of computer and internet technologies, texts, which were regarded as an important media for information exchange, had been used in storages and exchanged lots of thoughts of people. The characteristics of digital texts are easily copied and spread, that made the copyrights of the texts face a big challenge. Watermarking technology is a method for copyright protection. The embedding of watermarking shouldn't affect the usability of carrier media, while it can be extracted at will. In a sense, the watermarking is transparent of outsiders, but the detection means can be provided by the algorithm.

Su et al. [1] were particularly pessimistic: "Raw text, such as an ASCII text file or computer source code, cannot be watermarked because there is no 'perceptual headroom' in which to embed hidden information." Some researches have made changes to the orthography and layout in particular textual file formats [2, 3]. But such approach can be trivially defeated by file reformatting, or re-capturing text with optical text recognition. The first published account that found 'perceptual headroom' in the fabric of text [4] suggested that the substitution of equivalent words or constructions could hide bits of data without affecting meaning. Atallah et al. [5] describe a proof-of-concept watermarking implementation based on sentence transforms that they judge to change meaning 'slightly', though to what degree is not quantified. These last two approaches depend crucially on the quality of syntactic and semantic analysis. So these approaches are suffered from indeterminacy in synonym selection and the difficult problem for Natural Language Processing (NLP for short) to decide in what circumstances a particular transform is appropriate.

Zero-watermarking technology defers from traditional watermarking technologies. It did not change any part of the original carrier media. By picking up the characteristics of the carrier media, zero-watermarking is some transform result of these characteristics, and be registered at notarial office as copyright before publishing. Since the zero-watermarking does not modify any part or properties of the carrier, the imperceptibility is assured.[6] In this paper, we put forward a robustness text zero-watermarking based on part-of-speech (POS for short) tag sequence.

In the following section we examine some extant text zero-watermarking algorithm. In section III we introduce chaotic function, which as a building block of our algorithm. In section IV, we put forward a watermarking algorithm. Section V provides the results of the experiments. Section VI is conclusion.

II. STATE OF THE ART

[7] put forward a zero-watermarking algorithm based on English. In this algorithm, the text characters are firstly determined by the punctuations, and then by XOR the encrypted characters and the encrypted watermarking information. The result is the zero-watermarking. The robustness of this algorithm is low, since it is easy to replace the word before punctuations with synonym or combining two sentences by dropped the punctuations.

Chinese character components are used as watermark entry point [8], which selected by key. Then according to the selected Chinese character component, Chinese characters are extracted from the text. Lastly, watermarking characters selected under chaotic function control. The robustness is improved, since there are 580 Chinese character components. Without key, attacker can not found from which characters formed the zero-watermarking. Stochastic modify text will not destroyed zero-watermarking efficiently. But [9] making use of synonym substitution technology, defined two active attack methods: Sync-attack and Birthday-attack. Combining the two attack methods, [9] put forward an active attack algorithm. The result of experiment shows

that it wouldn't modify text massively but destroyed zero-watermarking easily. In order to ensure the imperceptibility, the expectation of attack success probability can be designated flexible.

III. DETERMINISTIC CHAOS

Chaotic is a state of disorder and irregularity. It describes many physical phenomena with complex behavior by simple laws.

Dynamical systems are such kind of systems that develop in time in a non-trivial manner. Deterministic chaos is irregular motion generated by nonlinear dynamical systems whose laws determine the time evolution of a state of the system from knowledge of its previous history.

Let A be a set. A function $f: A \to A$ is called chaotic on A if:

1) f has sensitive dependence on initial conditions.
2) f is topologically transitive.
3) Periodic points are dense in A.

A chaotic function $f: A \to A$ has sensitive dependence on initial conditions if there exists $\delta > 0$ such that, for any $x \in A$ and any neighborhood N of x, there exists $y \in N$ and $n \geq 0$ such that $\left| f^n(x) - f^n(y) \right| > \delta$. It mean that for each point x there is at least one point y in any neighborhood of it, which will eventually separate from x by a distance of at least δ after a certain number n of iterations of the function.

The quadratic family is such chaos functions. The functions of the quadratic family are defined as:

$$f_p(x) = px(1-x) \tag{1}$$

The following hold:

- $f_p(0) = f_p(1) = 0$ and $f_p(q_p) = (q_p) = q_p$ where $q_p = (p-1)/p$.
- $0 < q_p < 1$ if $p > 1$.

This means that there is at least one fixed point for each function of the family. Where parameter p controls the orbit of the map and $p \in [0, 4]$, $f_p \in [0, 1]$. For clear, we reformatted it in recursion version:

$$x_{n+1} = px_n(1-x_n) \qquad x_n \in [0,1] \quad p \in [0,4] \tag{1`}$$

The number of iterations is defined to be equal to the carrier characteristic. The above mentioned functions produce sequences of theoretically infinite period.

IV. THE TEXT ZERO-WATERMARKING

A. Conception of Zero-watermarking

Zero-watermarking technology defers from traditional watermarking technologies. It did not change any part of the original carrier media. By picking up the characteristics of the carrier media, zero-watermarking is some transform result of these characteristics and its length is conventional fixed beforehand. These characteristics should (1) represented the carrier media feature and (2) satisfied the collision resistance, which means that from different carriers, the probability that extracted zero-watermarking are same is very low. Right owner registered the zero-watermarking at notarial office as copyright before publishing. When suspect pirate, extracting zero-watermarking from the suspicious object. Comparing the extracting zero-watermarking and the registered zero-watermarking, if they are same, pirate is proved.

B. Part-of-speech tag sequence of texts

From linguistics, sentence pattern is represented by POS tag sequence. Sentence patterns are distinguished by quantity of POS tag and the permutation of POS tag sequence. Theoretically speaking, human can build up sentence as many as they want. So there are infinite sentence patterns, or POS tag sequence. For different text, their content and expression are different. their POS tag sequence is bound to different. We can draw the conclusion safely that POS tag sequence is unique characteristic of text.

But the whole POS tag sequence of text may be too long to be watermarking, that it also is not qualified for the fixed length requirement. On the other hand, if attacker can modified the POS tag sequence easily through performed meaning-preserving transformations of sentences of the text, such as adjunct movement, passivization and so on.

For robustness and efficiency, we adopt that, firstly, for English, there are 36 tags [14]. Three POS tag permutation build up 36^3 patterns of POS tag sequence, It means there are so many possible entry points. For attacker, it is hard to guess which pattern of POS tag sequence is the watermarking entry point. A special pattern is random picked up, and then we are combining all of these patterns appeared in the text into a sequence according to the pickup order, which we call it *characteristic part-of-speech sequence*. But it still has two shortages. Secondly, we random picked up some of the words from words sequence corresponding to *characteristic part-of-speech sequence*, which we call it *characteristic word sequence*. Lastly, some transformation is performed on the *characteristic word sequence*. The transformation result is the final zero-watermarking.

C. The Zero-watermarking Algorithm

Let T denotes the text, k denotes the key, w denotes zero-watermarking. Let s^3 be a pattern which POS tag sequence length is 3. Let S^3 be a set of all possible permutation of triple POS tags, which every element in S^3 is denoted by s_i^3.

The zero-watermarking algorithm described as follow:

Input: T and k;

Output: w

Begin

1. T is tagged by POS tagger and POS tag sequence s_n is the tagging result. n means the total word of the text is n.

2. $\forall s_i^3 \in S^3$, statistics frequency of $f(s_i^3)$.
3. An s_i^3 is random selected under the control of k, where $f(s_i^3) > a$.
4. $List(s_i)$ denoted combining all of words of s_i^3 patterns appeared in the text into a sequence.
5. Let *m* denotes the length of $List(s_i)$, where *m* is the number of iterations of the recursive function (1`). Where the parameter $x_0 = i * 0.00002$ and $p = random(k) * 4$ (*random* is a pseudo-random function return a value ranged from 0 to 1). Every recursion, the results of function (1`) are formed a chaotic sequence x_j (j = 1, 2, …, m).
6. Binarization result is obtained by function (2)

$$\zeta_j = \begin{cases} 1 & x_j \geq \sigma \\ 0 & x_j < \sigma \end{cases} \quad (j = 1,2,...,m) \qquad (2)$$

7. The $List(s_i)$ will be filtered by function (3).

$$m = \sum_j \zeta_j \wedge List_j(C_i) \quad (j = 1,2,...,m) \qquad (3)$$

Where Σ is a connection operator, which connect words into string and ∧ is union operator, if $\zeta_j = 1$ then the *j*th word will be added into m, else the *j*th word will be ignored. The filter out result string is called *watermarking word sequence*.

8. Return $w = SHA(m)$. Where *SHA* is hash function SHA-1.

End.

Then, the owner registers *w* into notarial office as copyright before publishing.

D. Validating Copyright

When pirate is suspicious, by performed the same algorithm on the suspicious text, watermarking is get. And comparing this watermarking and the registered watermarking, whether it is pirate is judged.

V. EXPERIMENTS AND RESULTS

In order to obtain our text data collection, we have employed the following procedure. First, we processed text from the Reuters Corpus [10] and 10000 stories are selected for experiments. We have used the Log-linear Part-Of-Speech Tagger [11] to obtain POS tag sequence of every story. Next, punctuation and currency symbols tags are filter out, since characters corresponding to these tags are irrelative semantics. Then, by getting rid of sentence boundaries, the number of possible permutation of sub-sequence of POS tag are more.

We then used the session IV algorithm to extract zero-watermarking of every story. Mass experiments show that when about 2~3% words of text are selected as *watermarking word sequence*, the collision resistance and robustness are best. In our experiments, in 10000 zero-watermarking, none of two zero-watermarking are identical.

A. Imperceptibility

Since the zero-watermarking does not modify any part or properties of the carrier, the imperceptibility is assured.

B. Security

According to Kerckhoffs principle, the attacker is unable to know which pattern of POS tag sequence is selected. Moreover, since chaotic function has sensitive dependence on initial conditions and unpredictability, the attacker can not fabricate the zero-watermarking without the key.

It is also infeasible that the attacker deduce the *characteristic word sequence* from zero-watermarking, since hash function has irreversibility. Furthermore, it is blind algorithm. When validating copyright, the original text is unnecessary.

C. Robustness

Compare to the changing formatting methods and character feature methods, reformatting, converting document type would not destroyed zero-watermarking, since the zero-watermarking is obtained from the text content, that is say, the word sequence of the text, which irrelative to text formatting or special font.

For malicious users, they also want to tamper with the text. There are two ways they could adopt. The first way is synonym substitution. Since the attacker did not know which sub-sequence POS tag was selected and which word was belong to *watermarking word sequence*, stochastic is inefficient. According to birthday attack, [9] have proved that if text totaled *n* words and *x* words were selected as *watermarking word sequence*, attacker has to substitute *n/x* words to destroy zero-watermarking safely. It is too many synonyms to find in a text.

For every story of the 10000 stories that we selected from Reuters Corpus, we have substituted synonyms stochastic. Figure 1 show the relationship between the percent of synonyms substituted (the X-axes) and the number of text that it's zero-watermarking was destroyed (the Y-axes). We see from the figure that synonym substitution attack is almost impossible success, since the POS sequence would not be changed by synonym substitution. So, the probability of collision resistance is every low.

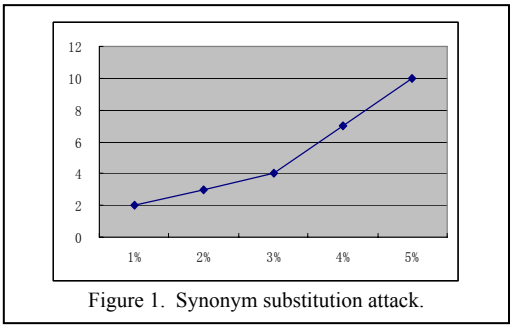


Figure 1. Synonym substitution attack.

The second way is making meaning-preserving transforms to plain text. For the same stories, we have transformed some sentences stochastic, which include adjunct movement, topicalization, extraposition, preposing, activization and passivization. Figure 2 show the number of text that it's zero-watermarking was

destroyed after transform sentences. The X-axes is the total words that affected by transformation, that is say the word may be added, dropped or moved. We can see from the figure that sentence transformation attack is much more efficient than synonym substitution attack, since the POS sequence will be changed by transformed sentences.

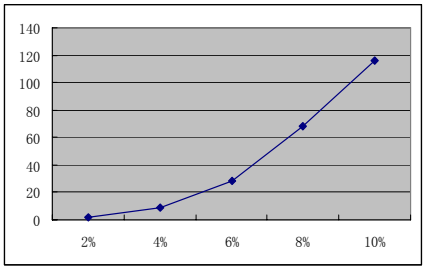


Figure 2. Sentence transformation attack.

He Lu et al. [9] put forward an efficient lexical active attack algorithm, which is not substituted synonym stochastic. In order to compare the robustness with Cheng's algorithm [8], we have selected 1000 stories from the People's Daily Corpus [12]. Next, we extracted the zero-watermarking by Cheng's algorithm and ours. Then we attacked these stories by the lexical active attack algorithm. Last, we reprogram the lexical active attack algorithm as syntactic active attack algorithm, since synonym substitution can not affected POS tag sequence. The results are showed in figure 3. The X-axes is the total words that affected by synonym substitution or sentence transformation. The Y-axes is the attack success probability.

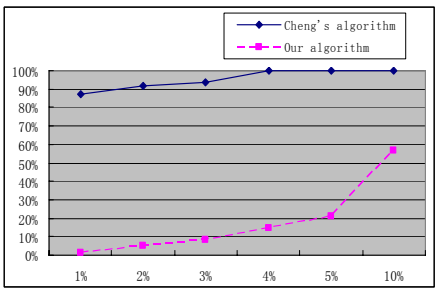


Figure 3. Compared robustness.

Since the second level Chinese POS tagset is 66 totally [13]. The number of pattern POS tag sequence is much more than the number of Chinese character component, which are only 580. So the robustness is improved dramatically.

VI. CONCLUSION

We presented and discussed a zero-watermarking based on POS tag sequence. Since there are so many possible permutation of POS tag sub-sequence and chaotic function has sensitive dependence on initial conditions or the key, the attacker impossible to build up a fake text or deduce the words of watermarking. Experiments result show that it is almost impossible destroyed the zero-watermarking by stochastic synonym substitution and sentence transform. Even under the well arranged attack, the robustness of our algorithm is much better than other's.

ACKNOWLEDGMENT

This paper is supported by Science Research Plan of Shaanxi Ministry of Education (09JK751), National University of Innovative Experimental Projects (091069720), Postgraduate Innovation Project of Northwest University (08YZZ32), Postgraduate Science Experimental Project of Northwest University (09YSY29).

REFERENCES

[1] J. K. Su, F. Hartung, and B. Girod, "Digital watermarking of text and image and and video documents," Computers and Graphics 22, pp. 687–695, 1971.

[2] J. Brassil, S. H. Low, N. F. Maxemchuk, and L. O'Gorman, "Electronic marking and identification techniques to discourage document copying," IEEE Journal on Selected Areas in Communications 13(8), pp. 1495–1504, 1995.

[3] W. Bailer and L. Rathner, "Linguistic information hiding." http://www.wbailer.com/wbstego, 2001.

[4] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," IBM Systems Journal 35(3-4), pp. 313–336, 1996.

[5] M. Atallah, V. Raskin, M. Crogan, C. Hempelmann, and K. Kerschbaum, "Natural language watermarking: Design and analysis and a proof-of-concept implementation," in Proceedings of Forth International Workshop on Information Hiding, 2001.

[6] Wen Quan, Sun TanFeng, Wang ShuXun. Concept and Application of Zero-Watermark, ACTA ELECTRONICA SINICA, 2003.2, Vol.31(2), pp. 214-216.

[7] Pan Li, Zou JianCheng. A Novel Zero-watermarking Algorithm Based on English Text Content. 12th China Youth Conference on Communications. Beijing, China, 2007. pp. 707-711.

[8] Cheng YuZhu, Sun XingMing, Huang HuaJun. Text zero-watermarking algorithm based on chaotic mapping, Computer Applications, 2005.12, Vol.25(12), pp. 2753-2754, 2758.

[9] He Lu, et al. A Lexical Active Attack Algorithm on Chaotic Text Zero-watermarking. Journal of Xi'an JiaoTong University, Submit.

[10] "Reuters corpus," http://about.reuters.com/researchandstandards/corpus/index.asp

[11] Stanford Log-linear Part-Of-Speech Tagger, http://nlp.stanford.edu/software/tagger.shtml

[12] the People's Daily Corpus, http://www.icl.pku.edu.cn/icl_groups/corpus/dwldform1.asp 2001.5.10.

[13] Liu Qun et al. Chinese POS Tag Set(ICTPOS3.0), http://www.nlp.org.cn/docs/docredirect.php?doc_id=993

[14] Mitchell P. Marcus et al. Building a Large Annotated Corpus of English: The Penn TreeBank. Computational Linguistics, 1993, Vol.19(2), 313-330. http://acl.ldc.upenn.edu/J/J93/J93-2004.pd

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 191-194

A Symmetric Image Encryption Scheme Based on Composite Chaotic Dispersed Dynamics System

Zhenzhen Lv[1], Lei Zhang[2], and Jiansheng Guo [3]
[1,2,3]Information Science and Technology Institute Zhengzhou, 45004, China
[1] E_summer@126.com

***Abstract*—By analyzing and researching several digital chaotic image encryption algorithms, this paper designs a symmetric image encryption scheme based on the three-dimensional Henon chaotic map and the general Cat chaotic map. In this scheme, the chaotic sequence of general Cat map is used as the initialization sequence to confuse the positions of image pixels; the three-dimensional Henon chaotic map can diffuse the image pixels; Then, repeat the confusing and diffusing process to increase the resistance to statistical and related attacks. Thorough experimental tests are carried out with detailed analysis, demonstrating the high security and fast encryption speed of the new scheme.**

Ⅰ. INTRODUCTION

During the development of network and multimedia technology, more and more images transmit over the Internet and through the wireless networks. The digital images become one of the most important information carrier which is helpful for people to communicate with each others. However, because of image's intrinsic features, such as bulk data capacity and high correlation among pixels, it is not suitable for practical image encryption, especially under the scenario of on-line communications. Therefore, people begin to explore a new image encryption pattern which is more efficient to hide the image information.

In the 1980s, chaos theory began to get involved in cryptography fields. Chaos sequence has some excellent characters, such as, the chaotic sequence is sensitive to the initialization; the output is approximate stochastic sequence; balanced statistical property; spreading the initial region over the entire phase space via iteration [1, 9]. Because of these, chaos theory has be used as a new direction for information and image encryption. At present, the image encryptions based on the chaotic sequence mainly research in low-dimensional (one-dimensional or two-dimensional) chaotic system. For example a image encryption based on Logistic map was widely used in [2, 3]. These image encryptions have some merits such as its form is simple and the time of producing the chaotic sequence is short and it is easy to release, but its keyspace is too small. In some condition, it can be attacked [5, 6, 7, 8, 10]. Consequently, the higher-dimensional chaotic system will become the new researching hot point.

A new approach is suggested in this paper for fast and secure image encryption. In order to fast confuse the relation among image pixels, a general Cat map is employed to shuffle the positions (and, if desired, grey values as well) of the pixels in the image. Meanwhile, to diffuse the relationship between cipher-image and plain-image, the three-dimensional Henon chaotic map is used as a diffusion processing among pixels. Thorough experimental tests are carried out with detailed analysis, demonstrating the high security and fast encryption speed of the new scheme.

Ⅱ. THE CHAOTIC MAP AND CONFUSING TRANSFORM

A. Henon map and general Cat map

Compared with the normal chaotic system, super chaotic system is more sensitive to initial conditions and parameters, and it has two positive Lyapunov index, and the output of system is more complex and stochastic. Therefore, it is more suitable for the security communication [3, 4].

The three-dimensional super Henon chaotic map can be defined as follow:

$$\begin{cases} x_{n+1} = a - y_n^2 - bz_n \\ y_{n+1} = x_n \\ z_{n+1} = y_n \end{cases}$$

where, $1.54 < a < 2$, $0 < |b| < 1$.

The general Cat map is a two-dimensional invertible chaotic map, and described by

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \bmod N$$

where, (x, y) is the the coordinate of plain-image pixels, (x', y') is the coordinate of the cipher-image pixels, N is the width or height of the image.One obtains a two-dimensional Cat map as follows:

$\begin{bmatrix} x' \\ y' \end{bmatrix} = A \begin{bmatrix} x \\ y \end{bmatrix} \bmod N$ It is easy to verify that when the det A=1, the Cat map is a bijection.

Set

$$A = \begin{bmatrix} 1 & a \\ b & ab+1 \end{bmatrix}$$

AP-PROC-CS-09CN005

the chaotic Cat map finally described as follows:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 1 & a \\ b & ab+1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \bmod N$$

where $x, y, x', y' \in \{0,1,2,\ldots,N-1\}$, and a, b are positive integers, and its converse map is as follows:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ab+1 & -a \\ -b & 1 \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix} \bmod N = A^{-1}. \begin{bmatrix} x' \\ y' \end{bmatrix} \bmod N$$

B. Confusing transform

Due to some intrinsic features of images, such as bulk data capacity and high correlation among pixels, if encrypt image directly, the image will have some regular character. And, it is very weak to the known-plain-text attack. Therefore, in this scheme confusing process will destroy the correlation of image pixels.

Take a size of $N \times N$ image for example. The confusing process is as follow:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = A \begin{bmatrix} x \\ y \end{bmatrix} \bmod N$$

The transform can confuse the positions of image pixels, but it can not change its pixel value. In order to keep the secret of image information, the scheme will diffuse the grey scale of image pixels through encryption.

III. IMAGE ENCRYPTION AND DECRYPT ION ALGORITM

In this paper, the chaotic symmetrical image encryption algorithm is based on the Henon map and the general Cat map. The key parameter is the two chaos map initial state (the precision is 64-bit). Set I_R is an original image, and its size is $M \times N$,and $(i, j, g(i, j))$ denotes this image, defined (i, j) is any image pixel coordinates, and $g(i, j)$ is this image pixel value $(1 \le i \le M, 1 \le j \le N)$. The complete image encryption scheme consists of nine steps, as shown in Figure 1.

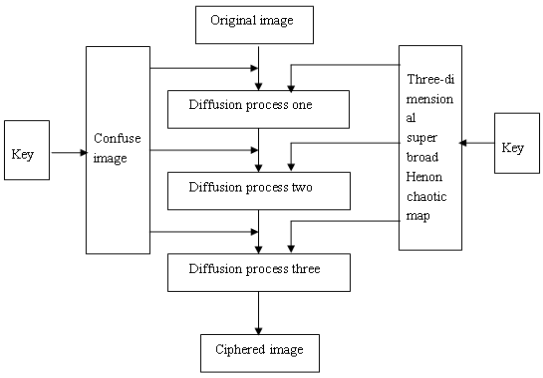


Figure 1. Block diagram of the image encryption

Step1 The initial parameters of three-dimensional super broad Henon chaotic map is key a, b, x_0, y_0, z_0. Iterate 200 times not to output, and from the 201st, output the iterative production chaos sequence, and it will generate three chaotic sequences $\{a_n\},\{b_n\},\{c_n\}$, where $n \in \{1,2,3,\ldots,N\}$.

Step2 Discretize the chaotic sequences. Set $\{x_n\}$, $\{y_n\}$, $\{z_n\}$, are key sequences, where $n \in \{1,2,3,\ldots,N\}$. The transform can be described as:

$$x_n = [a_n \times 2^{16}] \,\&\, \&0xff$$

$$y_n = [b_n \times 2^{16}] \,\&\, \&0xff$$

$$z_n = [c_n \times 2^{16}] \,\&\, \&0xff$$

where $[x]$ is adopt the integer of x.

Step3 The key chaotic Cat map is used to confuse the positions of image pixels. This process is the first time confusing image, by using the following formula:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = A \begin{bmatrix} x \\ y \end{bmatrix} \bmod 256$$

where det A=1.

Step4 Use the key sequence $\{x_n\}$ as the first time diffuse process, increasing the correlation of different pixels. The detail operate is as follow:

where $y \neq 0$, then

$$g'(x, y) = \left[\left(g(x, y) \oplus x_n \right) + g'\left(x, y-1 \right) \right] \bmod 256$$

where $y = 0$, then

$$g'(x, y) = \left[\left(g(x, y) \oplus x_n \right) + g'\left(x-1, 255 \right) \right] \bmod 256$$

especially,

$$g'(0,0) = \left[\left(g(0,0) \oplus x_1 \right) + g_0 \right] \bmod 256$$

Step5 Confuse the current image for the second time.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = A \begin{bmatrix} x \\ y \end{bmatrix} \bmod 256$$

where det A=1.

Step6 Diffuse the current image for the second time.

where $y \neq 0$, then

$$g'(x, y) = \left[\left(g(x, y) \oplus y_n \right) + g'\left(x, y-1 \right) \right] \bmod 256$$

where $y = 0$, then

$$g'(x, y) = \left[\left(g(x, y) \oplus y_n \right) + g'\left(x-1, 255 \right) \right] \bmod 256$$

especially,

$$g'(0,0) = \left[\left(g(0,0) \oplus y_1 \right) + g_0 \right] \bmod 256$$

Step7 Confuse the current image for third time, using the following formula:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = A \begin{bmatrix} x \\ y \end{bmatrix} \bmod 256$$

where det A=1.

Step8 The third diffusing process is as follow:

where $y \neq 0$, then

$$g'(x,y)=\left[\left(g(x,y)\oplus z_n\right)+g'\left(x,y-1\right)\right]\bmod 256$$

where $y=0$, then

$$g'(x,y)=\left[\left(g(x,y)\oplus z_n\right)+g'\left(x-1,255\right)\right]\bmod 256$$

especially

$$g'(0,0)=\left[\left(g(0,0)\oplus z_1\right)+g_0\right]\bmod 256$$

Step9 The output is the cipher image.

The decipher procedure is similar to that of the encipher process, with reverse operational sequences to those described from Step3 to Step8. Since both decipher and encipher procedures have similar structures, they have essentially the same algorithmic complexity and time consumption.

Ⅳ. THE ENCRYPTION EXAMPLE AND SECURITY ANALYSIS

A. The encryption example

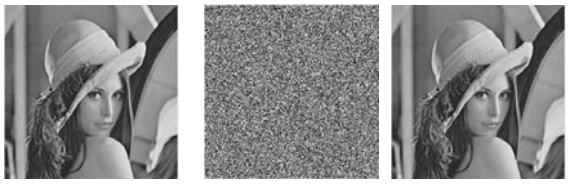
In order to confirm the algorithm's validity, the experiment has been taken. Set an image of size 256×256 and the initial key are:

$$a=1.98999999, b=0.0000899$$

$$x_0=1, y_0=2, z_0=3, g_0=0$$

$$A=\begin{bmatrix}1 & 1\\ 2 & 3\end{bmatrix}$$

Figure 2 shows the encryption example. Among them, (a) is the original image, (b) is cipher image, (c) is decipher image which using the right key. Through (b), one can see that the figure of plain-image pixels has be confused completely, and it is evidently different from the original image.



(a) (b) (c)

Figure 2. Test result

Through the experimental tests, show that this algorithm is valid and efficient, the encryption process and the decryption process can carry on smoothly.

B. The security analysis

A good encryption scheme should resist all kinds of known attacks, such as known-plain-text attack, cipher-text-only attack, statistical attack, differential attack, and various brute-force attacks. Some security analysis has been performed on the proposed image encryption scheme.

The key space. For the proposed image encryption algorithm, the key parameter is the initial parameters, including a,b,x_0,y_0,z_0,g_0 . Among them, the initial parameter a,b are 128-bit, the select space of x_0,y_0,z_0,g_0 are N (N is the largest grey scale in image). So the key space is $2^{128}\times N^4$, it is large enough to make brute-force attacks infeasible.

Key sensitivity test. A good image encryption algorithm should be sensitive to the cipher key; otherwise the opponent can break the cryptosystem by comparing two pairs of plain-text and cipher-text to discover some useful information. A key sensitivity test is performed as the following steps. First, a 256×256 image is encrypted using the key:

$$a=1.98999999, b=0.0000899$$

$$x_0=1, y_0=2, z_0=3, g_0=0$$

$$A=\begin{bmatrix}1 & 1\\ 2 & 3\end{bmatrix}$$

Then use different key to decrypt the ciphered image. Compared with the right key, the test keys have only 10^{-8} difference. Set the test keys are A and B, In A, the key parameter a is different; While in B the key parameter b is different. Using the two trivially modified keys to decrypt the cipher image.

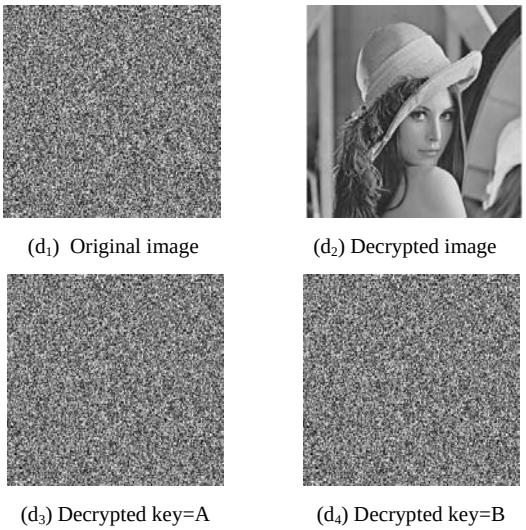
A: $a=1.98999998, b=0.0000899$

$x_0=1, y_0=2, z_0=3, g_0=0$

B: $a=1.98999999, b=0.0000898$

$x_0=1, y_0=2, z_0=3, g_0=0$

Figure 3 shows the key sensitive test results. Among them, (d_1) is the original encrypted image, (d_2) is decrypted image by right key, (d_3), (d_4) are decrypted image by the test keys. When the test key is slight different with the right key, the decryption process is completely failed. So, the algorithm can resist the divide-and-conquer attack.



(d_1) Original image (d_2) Decrypted image

(d_3) Decrypted key=A (d_4) Decrypted key=B

Figure 3. Key sensitive test result

Statistical analysis. Due to the continuity of image and the distribution of image pixels are not balanced. Statistical analysis has been performed on some proposed image encryption algorithms. the result demonstrates its superior confuse and diffuse properties which strongly resist statistical attacks. From the Figure 4, one can see that the histogram of the ciphered image is fairly uniform and is different from that of the original image.

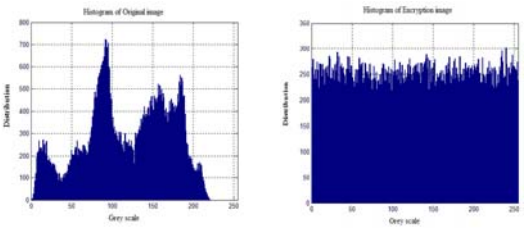


Figure 4. Histograms of the plain-image and the cipher-image

Ⅴ. CONCLUSION

The designed symmetric image encryption scheme employs the three-dimensional Henon chaotic map and general Cat map to process the original image independently. By confusing and diffusing transform, the positions and grey values of image pixels have been shuffled and increase its resistance to various attacks such as the statistical and differential attacks. The test and security analysis demonstrate the high security and fast speed of the new image encryption scheme.

However, designing a chaos image encryption scheme which is satisfied both high security and effectively needs is a difficult work. The algorithm of this paper is designed by certainly environment. Whether it exists some undiscovered methods to influence the security of this scheme, it needs to be researched.

REFERENCES

[1] Wang Shihong, Ye Weiping, Lu Huaping, “A Spatiotemporal-chaos-based Encryption Having Overall Properties Considerably Better than Advanced Encryption Standard”, http://arxiv.org, number: CD/0303026, 2003.

[2] Yuan Chun, Zhong Yuzhuo, Yang Shiqiang, “Composite Chaotic Pseudo-random Sequence Encryption Algorithm for Compressed Video”, Tsinghua Science and Technology, 2004, 9(2): pp. 234-241.

[3] Li Xiongjun， Peng Jianhua， Xv Nin， “A Image Encryption Algorithm Based on Two-dimensional Chaotic Sequence”, Journal of Image and Graphics， 2003， 8(10)：pp. 1172-1177.

[4] G Baier， M KLein, “Maximum Hyperchaos in Generalized Heron Maps”, Phys Lett A， 1990， 151（6/7）, pp. 281-284 .

[5] Kocarev L．Jakimoski G, “Pseudorandom Bits Generated by Chaotic Maps”, IEEE Transactions on Circuits and Systems-I： Fundamental Theory and Applications， 2003，50(1)：pp. 123-126．

[6] Guo Jiansheng, Zhang Cong’e, Jing Chenhui, “Attack with Know Image to a Symmetric Image Encryption Scheme”, Systems Engineering and Electronics, 2007, 29(3): pp. 341-345.

[7] Jing Chenhui. “Analysis of A Block Cipher Based on Chaos”, Engineering Science, 2001, 3(6): pp. 75-80.

[8] Jing Chenhui, Gao Heiying, “Analysis of Two Stream Ciphers Based on Chaos”, Acta Electronica Sinica, 2004, 32(7): pp. 1066-1070.

[9] Yuan Chun, Zhong Yuzhuo, Yang Shiqiang. Composite chaotic pseudo-random sequence encryption algorithm for compressed video. Tsinghua Science and Technology, 2004, 9(2): 234-241.

[10] Jing Chenhui. Analysis of A Block Cipher Based on Chaos, Engineering Science, 2001,3(6): 75-80

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 195-200

An Overview on Game Cheating and Its Counter-measures

Xiao Lan[1], YiChun Zhang[2], and Pin Xu[3]
[1,2,3]Digital Media Technology, Communication University of China, Beijing, China
Email: [1]broodblue@126.com, [2,3]{zhangyichun, xupin}@cuc.edu.cn

***Abstract*—Game cheating has become a serious problem embarrassing the long term development of computer game. However, there's no complete taxonomy survey on game cheating up to date yet. In this paper, we give the strict definition and complete taxonomy on game cheating. Our taxonomy is with cheating logic, and is classified by cheating motivation, cheating principal, cheating executants, cheating environment and the consequence. What's more, we give a classification and survey on relevant cheating countermeasure methods.**

***Index Terms*—computer game; online game; game cheating; cheating countermeasure**

I. INTRODUCTION

A. Background

Thanks to the booming development of computer and networking technologies, computer games have profoundly influenced our daily life. The games before 1990's are only within one player mode that the opponent is computer AI. We call them console games. In recent years, online games have become prevalent in the game production industry and they are popular with youths.

However, for all the game players, their fulfillment is to win. Some players want to win without corresponding efforts and with unfair advantages over opponents, so game cheating emerges. In order for fair players own sake and game development, it's necessary for us to look into game cheating behaviors, summarize something in common and work out corresponding countermeasures at last. However, there's no systematic and integrated overview on game cheating, and it limits the further investigation. As far to our knowledge, there're three problems need to be solved when someone wants to game cheating summarization.

The diversity of cheating behavior: As the game style differs from one to another (such as 1st-person shooters FPS, massively multiplayer on-line role-playing games MMORPG, and peer-to-peer games P2P[1]), corresponding cheating techniques are complicated and multiplex. Furthermore, new cheating manners are emerging with the development of game.

The insecurity of virtual world: The rules in virtual are different from those in real world. Game server cannot guarantee all personal data of the players in the virtual world are reliable and secure. When a cheater join a game using an untruthful personal data, it's possible for him doing whatever he wants without supervision.

The difference between game cheating and traditional security attack: Game cheating is an emerging security problem. Comparing to traditional security-related threat, game cheating differs in motivation, target, principle, consequence and so on. For example, cheater uses game-bot (malware in moral view) intentionally, but in convention security scenario, malware executes violating user's intention.

B. Related work

As game cheating is evolving rapidly, it is important that make sure what game cheating means to players and game corporations. If people reach a consensus on game cheating, it's helpful keeping the virtual game world orderly. Up to now, many researchers have attempted giving a precise definition of game cheating or constructing a framework to classify game cheating.

Definition: In order to distinguish smart play from cheating, Yan[2] define the cheating as" Any behavior that a player may use to get an unfair advantage, or achieve a target that he is not supposed to be." But by this definition, it's difficult to judge player's actions as cheating because how to define "unfair advantage" is ambiguous. As an improvement, a more sufficient definition was presented by Yan[3] as "any behavior that a player uses to gain an advantage or achieve a target in an online game is cheating if, according to the game rules or at the discretion of the game operator the advantage is unfair to his peer players or the target is one that he is not supposed to achieve". The improved definition focuses on the online cheating only, console game is excluded. So we define game cheating with minor modification as "any behavior that a player uses to gain an advantage out of the context of the game rules' permission or achieve a target that he is not supposed to be is game cheating". Additionally, the rules should contain not only the rules in virtual world also the rules related to the game in real world.

Taxonomy: Pritchard [4] listed the game cheating manners that have occurred in various games. And he proposed a six-category which comprised of reflex augmentation, authoritative clients, information exposure, compromised servers, bugs and design loopholes, and environmental weaknesses. Reflex augmentation places extra emphaises on producing superior results from player's reaction. For example, aiming proxy in FPS games is a form of this type. The main form of authoritative clients cheating is modifying the game data such as modifying memory data to change the game properties and altering the network communication data packets to effect the result of certain game events. Information exposure is different from authoritative clients, cheating with information exposure does not modify game data directly but just expose them to cheater,

and cheater gains game knowledge advantage over opponents. For example, the cheater catches the data packets from network, and then analyzes them to obtain the hidden information such as the location of online opponents. The cheating with compromised servers occurs at game supporting server. The cheater first attacks the game servers using some familiar hacking techniques, and then modifies the server data or gets the information of other players. As the implemenataion of game architecture and game rules is complex and imperfect, exploiting bugs and design loopholes become a popular way of cheating. Environmental weaknesses is "something of a catchall for exploitable problems a game may have on particular hardware or operating conditions[4]". For example in some games, cheater uses advanced display card which is of high brightness to show the dark places. But this categorization doesno't cover all types of cheating manners. As cheating behaviors are evolving, many emerging cheating behaviors cannot fit into any of these categories. For example, some behaviors such as cheating collusion, operator cheating and client hacking are not included.

In [2], game cheating has been extended into 11 commom cheating classes. Since new cheating forms are added, 4 more cheating were added in [5]. Meanwhile, according a three dimensional taxonomy, online cheating was classified by "the underlying vulnerability (what is exploited), the cheating consequence (what type of failure can be caused), and the cheating principal (who is cheating)." This framework is helpful for cheating understanding, but their work doesn't cover the counter-measures.

In order to make better understanding emerging research area of game cheating and boost anti-cheating improvement, we propose a novel classification of game cheating depending on cheating motivation, cheating principal, cheating location and its consequence in a more logical way. What's more, we do an overview of current anti-cheating techniques and point out their relations to our cheating taxonomy.

The novel taxonomy will be presented in Section II. We present relevant anti-cheating methods and potential improvement in Section III. And conclusion is given in Section IV.

II. NOVEL TAXONOMY OF GAME CHEATING

We classify the game cheating in a logical order. The relationship between cheating behavior and our classification is shown in TABLE I. The more detailed illustrations are as follows.

A. Motivation of the cheating

In console games: The motivation is trivial. It is to defeat the computer AI opponent easier and faster.

In online games: There're three main goals as follows:

Defeat the competitor or avoid failure: In most P2P and FPS games, players cheat to augment the probability of winning.

Get higher game level: Higher level means stronger power, access to advanced plot and usufruct of better items. For example, due to the amount of time needed to gain experience and items in massively multiplayer online role-playing games, some people have resorted to cheating to take a short cut to the higher levels of the game.

Obtain economic interest: Virtual characters and items in online games can be traded for real money. A cheater might offer less and get more virtual advanced items by some kind of cheatings, so he can turn the virtual stuff into real money[6].

B. Approach of cheating

Modify game software or data: Running online game needs to install game client software and connect to Internet. So cheater can modify the software directly such as removing validating routines, modifying configuration parameters, rewriting some parts of game software and so on[7]. Also, when game software is running in the memory, cheater can modify the memory data so as to influence the running game because of most game attributes are stored in the memory. When a game is running within local memory, cheater can look for the critical variables and then change them. Game sensitive data can be deteced not only in local memory but in the packets transfering via the networks and in server's database. Game related packets include commands and sensitive information so that cheater can insert, delete or modify the packets to illegally obtain competition advantage. In the online fighting action game, cheater may modify the data packets to strengthen destructive power of the cheater's role[17]. The data stored in server's database usually contain player's private information and game state, they can be modified by game operator or administrator which forms a new kind of cheating.

A third-party tool but not modifying games: The cheating tools may be some softwares, number of small programs, game bots, or intelligent computers. Instead of modifying the data directly, some tools collect the information from the memory and networking packets, analyze these information, and form a clear report on game state. For example, cheater can determine the competitor's location by analyzing packets which contains game sound information. For another instance, in the chess tournament online, if a cheater uses a computer to predict opponent's behavior and design a best policy, he will gain huge advantage on the next move easily (just follow computer's prediction). Via modifying the graphic drive, cheater can make the walls transparent or help themselves accuratly aim. Game bots are common in most online games. For example, players use bots to perform repetitive tasks to save a great deal of time. Game bots can shoot more accurately in FPS. In order to gain high scores and the reputation, cheaters ues bots to join in the game instead themselves[16]. In addition, cheater can also deny other players' service with a network attacking tool. When networking condition is harsh, players can not play the game normally. The cheater may can use some flooding attacks, blocks the communication between server and competitor, delays opponents' response. Finally, opponents are likely to be kicked out of the game[10] according to current game

rules. Victim player in car racing online games will loose continuity of driving if he suffers from cheater's flooding attack.

Collude with other players or game operator: People can collude with other players to gain unfair advantage over their honest opponents in online games[8]. In online bridge games or poker games, cheaters illicitly exchange card information over the other communication tools so that they can grasp the information of the honest competitors. Another typical example is "win-trading" in which cooperative players lost to the other alternatively in order to raise cheaters' victory numbers[14]. Furthermore, cheaters can collude with the game operators who have access to modify the game database. Collude operators are able to create a strong role for cheaters.

Exploit bug or loophole: Not only bugs of game design are expoited, but also the rules loophole[8]. This kind of cheating is resulted from system design flaws and imperfect game rules or policies. A good example is a so-called "camping" behavior in online war games[2]. The cheater stay in a corner where the other players can't shoot him easily. Then he is only to wait for the time running out or killing the enemy who is passing the place. The "corner" is one loophole of the game design. Moreover, making use of rule loopholes, cheater may advoid failure. When the player will lose the game, he may escape the game or make himself disconnecting from the game. So in his record, there will be less failures.

Time cheating: A cheating player can delay his own move until he knows all the opponents' moves. So he can obtain more time to react to a game event than a honest player. Time cheating involves message exchange, game state update, network and computational latency and so on [9].

Steal other player's ID/Password: ID or password is the key for player accessing to the game system[7]. It is also used for keeping his account alive and doing business in virtual world. Therefore, driven by the economic interest, the cheater tries everything available to steal other player's ID and password. Stealing can be approached by attacking server's database or accessing to other's personal PC with common hacking techniques.

C. number of cheater

Single cheater; Single cheater can cheat either in single-player or multi-player online games. Single cheater may be a game player or a game operator[5]. Almost single cheater cheating are executed by single player except modifying servers' database is conducted by game operator.

Multiple cheaters: This cheating usually happens in multi-player online game. The cheaters can collude with other players or game administrator. Collusion among players results in game victory, but cheating players cannot gain extra benefits. However, if the collusion involves game administrator, things go even worse. Cheaters are not only with game victory, it's possible for they to obtain a stronger role or advanced items. Especailly, cheaters is able to steal the other players' ID/Password, and turn it into the economic interest (real money).

D. Location of cheating

In modern online game world, there are three types game based on networking infrastructure: distributed game, centralized game, and client/server game. So we

TABLE I. TAXONOMY AND RELATIONSHIP OF GAME CHEATING BEHAVIORS

		Approach					
		Modify game software or data	A third-party tool but not modifying games	Collude with other players or game operator	Exploit bug or loophole	Time cheating	Steal other player's ID/Password
Motivation	Defeat the competitor or avoid failure	√	√	√	√	√	
	Get higher game level	√	√	√	√		
	Obtain economic interest	√		√			√
Cheater number	Single player	√	√		√	√	√
	Game operator	√					√
	Multiple players			√			
	Player with operator			√			√
Location	Client	√	√		√		√
	Server			√			√
	Network		√			√	
	Real world			√			
Consequence	Game enjoyment losing	√	√	√	√	√	
	Threaten security of other's PC		√				√
	Economic losses	√		√			√

divide the cheating scene into four classes.

Act at client: Online game software is installed at the client side. The cheater use assistant tool which was running on client side. Frankly speaking, game software implementation is imperfect, the player will find out the bugs and loopoholes eventually. Cheating behaviors such as modifying game software or memory , using a third-party tool and exploiting bugs or loopholes is inevitablely at client's computer.

Act at servers: Server's database stores Information of all players, including ID/Password, role information and so on. In special business platform, player can trade the game characters or items into real money. If cheater attack the server, he is able to obtain other players' information and change game role attributes. Usually, cheaters achieve the goal using hacking tools. However, the game operator who has the privilege to manage the game is able to cheat easiy and change anyone's information at his will. So it happens sometime that cheater bribes operator for game benefits.

Act in Network: Network is the only communication media between clients and servers, player and his competitor. It is important to build effective networking enviornment supporting online game. Cheaters modify the data flow between client and server, and influence the game result by changing the game commands or actions. Cheaters also can analyze the data to expose the hidden information. For example, time cheating involves data exchange and data update.

It helps cheaters obtain more reaction time. Another example is flooding attack. Cheater blocks competitor's communication so that victim will result in incorrect game reaction.

Act in Real world: Cheating happens not only in virtual world but in real world as well. Game collusion is a traditional cheating manner which happens in the real world. Colluders illicitly exchange the game information so that they gain unfair advantage over their honest opponents, and it's beyond the virtual game rules.

E. consequence of cheating

Different cheating behaviors may cause different consequences. We list four main consequences as follows:

Harmless: In console games, cheating is only the tool with which the player can finish the game easier and faster. These cheatings doesn't harm other pepole.

Game enjoyment losing: The joy of the game exists in the fierce competition. But cheating destroys the balance among players. Then game enjoyment for whole is losing. The more common cheating emerges in a game , the more fair players lose their interests .

Threaten the security of other players' PC: Hacking techniques are used in game cheating for the purpose to steal other players' IDs/Passwords. This means that players' PC will in danger when cheating happens .

Economic losses: IDs and Passwords losing may result in the accounts and virtual items losing. For example, player can reach high level or get advanced items more easily via cheating. Reduced gaming time will cut down the income of the game service provider[18].

III. ANTI-CHEATING SOLUTIONS

Using the cheating classification above, we can classify our counter-measures more explicitly, the following is relevant anti-cheating methods. The taxonomy of anti-cheating solution depending on cheating classification is presented in TABLE II.

A. Education and punishment

According to different cheating motivations, game service provider can make different education and punishment policies at the same time respectively. Education is that let players realize the moral and economic risks on cheating. Meanwhile, the punishment policy against cheater is getting rigorous. Once the player is found cheating, his ID/Password will be banned. So he will lose all including his game role, the game record, the items even his account.

B. Anti-cheating methods

Software and data protection: Encryption is used to prevent data from modification. Encryption can encrypt critical information in memory and in the transferring packets so that the attackers can't recognize the location of the elements and change them. TRM (Tamper Resistant Module) is one effective method to prevent the software cracking. TRM can verify the integrity of software whether the modification occurs in it [7]. Binaries protection is also an effective solution in the long term to prevent modification. Binaries protection [12] introduces dynamic mobile agents that an original agent is periodically downloaded and executed. So it is difficult for the cheater to break it due to the download code is always new. In [19], such a technique is also presented. "Mobile Guard" is used to ensure "the integrity of the protection mechanisms the solution does not statically embed them into the game-client". At the same time, Randomly Created Checksum Algorithms (RCCAs) enforce Mobile Guard to be executed. The limitation of this method is that it only prolongs protection time but is not a complete protection all the time.

Detection: For the data analyzing, encryption and binaries protection is also applied because the information comes from the memory data and network packets. For the third party programs, servers can detect them by scanning the memory of the client computer. Cheating detection also could be provided by honest players' prosecution. If they find the people who act abnormally, they can report to game administrator. The third method of detection takes place on the servers. The servers record players' actions and analyze them. By digging into the log, servers can find out the cheater [15]. Most of the detection methods are bots detection ones. Golle and Ducheneaut [20] propose two approaches to prevent bots with the CAPTCHA tests. This test is effective but it spoils the game continuity of the honest player. Except above method, we can use passive detection method. In [16], comparing real-life traces with the avatar's movement controlled by the players directly, researchers propose a trajectory-based approach. In [21], researchers show the traffic differences between those

generated by bots and by human players in various aspects. But the two works focus on one side only. The former assumes the bots controlled by the player directly, and in the later, the bots work as a standalone client.

Random assign and operator management:

Deal with multiple-players cheating: When the player joins in the game, system will distribute him into the room randomly. But this solution separates players who want to play with the partner he is familiar with. Therefore, current game usually provides two different running models: one is random, the other is unlimited.

Deal with player with game operator cheating: To prevent the information leaking out, it can be solved by reducing the power of game operator, building more rigorous punishment policy and logging the game operating events.

Enough test and timely update: It is impossible that one game is perfect without bug or loophole. Cheating cannot be avoided but can be reduced. Before the game shipment, the game developer need take more time to scrutinize the game. And anytime they find the bug or loophole, they need to make the update package timely.

Cheat controlled protocol: Cheater using time cheating who controls the communication message has additional time to react to honest player's action. In [11], the paper presents a protocol that can be used to control cheating in reaction time based message ordering schemes.

Protection of information and PC: The attacking target of cheaters may be the servers or the player's computer. So installing a personal firewall is necessary for client computer to protect personal information and ensure PC security. At the same time, it is vital to improve the central server's security, and robust mechanisms should be introduced identifying the ID/Password. Also, detection mechanism is necessary to find out the cheaters.

C. Proof of identity for player

At the present time, game service provider still lack a systemic authentication scheme to judge the validity of players' identity. The cheaters can use false personal data to login in the game and then destroy games. If their passwords are banned, they can register another one. So strengthen the identity management is extremely urgent.

D. Protocol-level solution

When cheating happens on network level, using anti-cheating protocol is the mainstream solution [12]. The protocols are created to restrict the time or mode of message transferring through network. For example, the protocol for time cheating judges the cheating behavior by transmission time of all kinds message. In [13], secure protocol based on public key cryptography is presented to detect cheating on P2P online games.

E. Risk management

Game company should build a special department dealing with prosecution reports from the players who suffered from cheating. ID/password should not be the only approach accessing to user's account. Digital certificate technique can be introduced. The account will be more secure with the combination of password and certificate.

TABLE II. TAXONOMY OF ANTI-CHEATING SOLUTION

Cheating		Anti-cheating Solutions
Motivation		Education and punishment
Approach	Modify game software or data	Software and data protection
	A third-party tool but not modifying games	Detection
	Collude with other players or game operator	Random assign and operator management
	Exploit bug or loophole	Enough test and timely update
	Time cheating	Cheat controlled protocol:
	Steal other player's ID/Password	Protection of information and PC
Cheater		Proof of identity for player
Location: in network		Protocol-level solution
Consequence		Risk management

IV. CONCLUSION

In this paper, a novel classification of game cheating is presented by cheating motivation, cheating approach, cheater number, cheating location and cheating consequence. Combining with reviewing the earlier research works, we get a more systematic and structured cognition on game cheating. The introduction about cheating motivation and cheater can help the service providers to find out who is with higher probability to be the cheater.

Nowadays, computer games have become one of most common computer applications in our daily life. Games do not only affect our entertainment life but change industry provision chain as well. The number of crime that caused by game cheating are continually raising [14]. But game security is still an emerging research area. Many existing problems in game cheating remain unsolved. People should pay more attention on game cheating and related security development.

ACKNOWLEDGMENT

This work is supported by "211 project" and "382 project" of Communication University of China.

REFERENCES

[1] D. Saha, S. Sahu, and A. Shaikh, "A Service Platform for On-Line Games," Proceedings of the 2nd workshop on Network and system support for games table of contents, *Redwood City, California,* pp. 180–184, 2003.

[2] J .Yan and H.J. Choi, "Security Issues in Online Games", *The Electronic Library, MCB, UP, Ltd*,Vol. 20, No.2, pp. 125-133, 2002.

[3] J.Yan, "Security Design in Online Games", *in Proc. of the 19th Annual Computer Security Applications Conference, IEEE Computer Society,* New York, pp.286-295, December, 2003.

[4] M.Pritchard, "How to Hurt the Hackers: The Scoop on Internet Cheating and How You Can Combat It", *Information Security Bulletin*, pp.33, February 2001.

[5] J.Yan and B.Rendell,"A systematic classification of cheating in online games", *NetGames 05, Hawthorne,* New York, USA,ACM, pp.1-9, October 10–11, 2005.

[6] J. Zetterström,"a legal analysis of cheating in online multiplayer games", Göteborg University, Master Thesis, March 2005.

[7] J. Ki, J. H. Cheon, J-Uk Kang, D. Kim,"Taxonomy on Online Game Securitty", *the Electronic Library,* Vol. 22, No.1, pp.65-73, 2004.

[8] P. J. Brooke, R. F. Paige, J. A. Clark, S. Stepney," Playing the game: cheating, loopholes, and indentit", *SIGCAS Comput*. Soc., Vol. 34, No. 2. September , 2004

[9] B.D.Chen and M.Maheswaran," A cheat controlled protocol for centralized online multiplayer games", *Proceedings of 3rd ACM SIGCOMM workshop on Network and system support for games,* Portland, Oregon, USA, ACM, pp. 139-143, 2004.

[10] H.B.-L. Duh and V.H.H. Chen, "Cheating behaviors in online gameing", *Online Communities and Social Computing,* Springer Berlin / Heidelberg, vol. 5621/2009, pp. 567-573, 2009

[11] J. Hu and F. Zambetta, "Security issues in massive online games", *Security and Communication Networks,* Vol. 1, No. 1, pp.83-92, 2008.

[12] S. Bernard, M.G. Potop-Butucaru and S. Tixecuil, "Cheats in online video games: detection, analysis, and countermeasures", [2008], http://www.thlab.net/old/rescom2008/posters/Samuel_Bernard.pdf

[13] H. Yoshimoto, R. Shigetomi and H. Imai, "How to protect peer-to-peer online games from cheats", *Proceedings of the Symposium on Information Theory and Its Applications*, Vol. 27, No.1, pp.315-318, 2004.

[14] R. Joshi, "Cheating and virtual crimes in massively multiplayer online games", *technical report*, Roal Holloway, University of London, January 2008.

[15] K. Warns, "Cheating Detection and Prevention in Massive Multiplayer Online Role Playing Games", *The Seventh Annual Winona Computer Science Undergraduate Research Symposium*, Winona, MN, April 2007.

[16] K. Chen, A. Liao, H. K. Pao and H. Chu, " Game Bot Detection Based on Avatar Trajectory", *Lecture Notes In Computer Science*, Vol. 5309 , Pittsburgh, pp.94-105, 2008.

[17] P. Laurens, R.F. Paige, P.J. Brooke and H. Chivers, "A Novel Approach to the Detection of Cheating in Multiplayer Online Games", *Proceedings of the 12th IEEE International Conference on Engineering Complex Computer Systems,* Auckland, pp.97-106, 2007.

[18] D. Pelland, "Hackers, Cheaters Threaten Online Games' Business Model",March 3,2005, http://www.kpmginsiders.com/display_analysis.asp?cs_id=126855

[19] C. Monch, G. Grimen, and R. Midtstraum, "Protecting online games against cheating", *in NetGames 06: Proceedings of 5th ACM SIGCOMM workshop on Network and system support for games*, New York , USA, ACM, pp.20, 2006.

[20] P. Golle, N. Ducheneaut, "Preventing bots from playing online games", *Computers in Entertainment*, Vol.3(3), New York, ACM, pp.3, July 2005.

[21] K.T. Chen, J.W. Jiang, P. Huang, H.H. Chu, C.L. Lei, W.C. Chen, "Identifying MMORPG bots: A traffic analysis approach.", *Proceedings of the 2006 ACM SIGCHI international conference on Advances in computer entertainment technology,* Vol.266, No.4, Los Angeles, USA,ACM, pp.20-23, June 2006.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 201-204

Type II Composed Fuzzy Measure of L-measure and Delta-measure

HsiangChuan Liu[1], DerBang Wu[2, 3], WeiSung Chen[4], HsienChang Tsai[5], YuDu Jheng[6], and TianWei Sheu[2]
[1] Department of Bioinformatics, Asia University, Taichung County, Taiwan
Email: lhc@asia.edu.tw
[2]Graduate Institute of Educational Measurement and Statistics, Taichung University, Taichung, Taiwan
[3]Department of Mathematics Education, Taichung University, Taichung, Taiwan
[4]Department of Computer Science and Information Engineering, Asia University, Taichung County, Taiwan
[5]Department of Biology, Changhua University of Education, Changhua, Taiwan
[6]Department of Education, Taichung University, Taichung, Taiwan
Email: wudb@hotmail.com, james@asia.edu.tw, bihft@cc.ncue.edu.tw, doora0622@yahoo.com.tw, sheu@mail.ntcu.edu.tw

***Abstract*—The well known fuzzy measures, λ-measure and P-measure, have only one formulaic solution. Two multivalent fuzzy measures with infinitely many solutions were proposed by our previous works, called L-measure and *δ*-measure, but the former does not include the additive measure as the latter and the latter has not so many measure solutions as the former. Due to the above drawbacks, an improved fuzzy measure composed of L-measure and half of *δ*-measure, denoted L(*δ*)-measure, was proposed by our other previous work. In this paper, a further improved fuzzy measure composed of L-measure and the whole of *δ*-measure, called Type II L(*δ*)-measure is proposed. For evaluating the Choquet integral regression models with the new fuzzy measure and other different ones, a real data experiment by using a 5-fold cross-validation mean square error (MSE) is conducted. The performances of Choquet integral regression models with fuzzy measure based on Type II L(*δ*)-measure, L(*δ*)-measure, L-measure, *δ*-measure, λ-measure, and P-measure, respectively, a ridge regression model, and a multiple linear regression model are compared. Experimental result shows that the Choquet integral regression models with respect to Type II L(*δ*)-measure outperforms others forecasting models.**

***Index Terms*—Lambda-measure, P-measure, Delta-measure, composed fuzzy measure, Choquet integral**

I. Introduction

When there are interactions among independent variables, traditional multiple linear regression models do not perform well enough. The traditional improved methods exploited ridge regression models [1]. Recently, the Choquet integral regression models [7-13] based on some single or compounded fuzzy measures [2-5, 7-12] were used to improve this situation. The well-known fuzzy measures, λ-measure [2-4] and P-measure [5] have only one formulaic solution of fuzzy measure, the former is not a closed form, and the latter is not sensitive enough. Two multivalent fuzzy measures with infinitely many solutions were proposed by our previous works, called L-measure [8-9] and *δ*-measure [10], but L-measure does not include the additive measure and *δ*-measure has not so many measure solutions as L-measure. Due to the above drawbacks, an improved fuzzy measure composed of L-measure and the first half of *δ*-measure, denoted L_δ-measure, was proposed by our other previous work [11].

In this paper, a further improved fuzzy measure composed of L-measure and the whole of *δ*-measure, called Type II L_δ-measure is proposed. This new fuzzy measure is more sensitive than L_δ-measure. For evaluating the Choquet integral regression models with our proposed fuzzy measure and other different ones, a real data experiment by using a 5-fold cross-validation mean square error (MSE) is conducted. The performances of Choquet integral regression models with fuzzy measure based Type II L_δ-measure, L_δ-measure, L-measure, *δ*-measure, λ-measure, and P-measure, respectively, a ridge regression model, and a multiple linear regression model are compared.

II. Fuzzy Measures

The two well known fuzzy measures, the λ-measure proposed by Sugeno in 1974, and P-measure proposed by Zadah in 1978, are concisely introduced as follows.

A. Axioms of Fuzzy Measures

Definition 1: fuzzy measure [2-4]

A fuzzy measure μ on a finite set X is a set function $\mu: 2^X \to [0,1]$ satisfying the following axioms:

$$1)\ \mu(\phi)=0,\ \mu(X)=1 \text{ (Boundary conditions)} \quad (1)$$

$$2)\ A \subseteq B \Rightarrow \mu(A) \le \mu(B) \text{ (monotonicity)} \quad (2)$$

B. Singleton Measures

Definition 2: singleton measure [2-7]

A singleton measure of a fuzzy measure μ on a finite set X is a function $s: X \to [0,1]$ satisfying:

$$s(x)=\mu(\{x\}),\ x \in X \quad (3)$$

$s(x)$ is called the fuzzy density of singleton x.

C. λ-measure

Definition 3: λ-measure [3]

For a given singleton measures s, λ-measure, g_λ, is a fuzzy measure on a finite set X, satisfying:

$$\begin{aligned} &A,B\in 2^X,\ A\cap B=\phi,\ A\cup B\neq X \\ &\Rightarrow g_\lambda\left(A\cup B\right) \\ &\quad =g_\lambda\left(A\right)+g_\lambda\left(B\right)+\lambda g_\lambda\left(A\right)g_\lambda\left(B\right) \end{aligned} \tag{4}$$

$$\prod_{i=1}^{n}\left[1+\lambda s\left(x_i\right)\right]=\lambda+1>0,\ s\left(x_i\right)=g_\lambda\left(\left\{x_i\right\}\right) \tag{5}$$

Note that λ-measure has a unique solution without closed form, and if $\sum_{x\in X}s(x)=1$ then λ-measure is just the additive measure.

D. P-measure

Definition 4: P-measure [5]

For given a singleton measures s, P-measure, g_P, is a fuzzy measure on a finite set X, satisfying:

$$\begin{aligned} &{}^{\forall}A\in 2^X \\ &\Rightarrow g_P\left(A\right)=\max_{x\in A}\left\{s\left(x\right)\right\}=\max_{x\in A}\left\{g_P\left(\left\{x\right\}\right)\right\} \end{aligned} \tag{6}$$

III. THREE MULTIVALENT FUZZY MEASURES

Both of λ-measure and P-measure have only one formulaic solution of fuzzy measure, three multivalent fuzzy measures were proposed by our previous works as follows.

A. L-measure

Definition 5: L-measure [8, 9]

For a given singleton measure $s\left(x\right)$, L-measure with determine coefficient $L\in\left[0,\infty\right)$ on X, is a multivalent fuzzy measure with determine coefficient $L\in\left[0,\infty\right)$ on X, satisfying:

1) $g_L\left(\phi\right)=0, g_L\left(X\right)=1$ (7)

2) ${}^{\forall}A\subset X, A\neq X\Rightarrow$

$$g_L(A)=\max_{x\in A}\left[s\left(x\right)\right]+\frac{\left(\left|A\right|-1\right)L\sum_{x\in A}s\left(x\right)\left[1-\max_{x\in A}\left[s\left(x\right)\right]\right]}{\left[\left|X\right|-\left|A\right|+\left(\left|A\right|-1\right)L\right]\sum_{x\in X}s\left(x\right)} \tag{8}$$

Note that L-measure has infinitely many solutions of fuzzy measures, for each $L\in[0,\infty)$, but it is not additive measure.

B. δ-measure [10]

Since L-measure does not include the additive measure, an improved multivalent fuzzy measure, called δ-measure was proposed by our previous work as following definition.

Definition 6: δ-measure [10, 11]

For given singleton measures $s\left(x\right)$, a δ-measure is a multivalent fuzzy measure with determine coefficient $\delta\in\left[-1,1\right]$ on a finite set X, $\left|X\right|=n$, satisfying:

1) $\sum_{x\in X}s(x)=1$ (9)

2) $g_\delta\left(\phi\right)=0, g_\delta\left(X\right)=1, g_\delta\left(\left\{x\right\}\right)=s\left(x\right),\ \forall x\in X$ (10)

3) $\forall A\subset X, 1<\left|A\right|<\left|X\right|\Rightarrow$

$$g_\delta\left(A\right)=\begin{cases}\max_{x\in A}s\left(x\right) & if\ \delta=-1\\ \dfrac{\left[1+\delta\max_{x\in A}s\left(x\right)\right]\left(1+\delta\right)\sum_{x\in A}s(x)}{1+\delta\sum_{x\in A}s(x)}-\delta\max_{x\in A}s\left(x\right) & if\ \delta\in\left(-1,1\right]\end{cases} \tag{11}$$

Note that δ-measure has infinitely many solutions of fuzzy measures, for each $\delta\in\left[-1,1\right]$, it is not so many solutions as L-measure, but it includes the additive measure.

C. L_δ-measure

Though δ-measure includes the additive measure, but it has not so many measure solutions as L-measure, therefore, an improved multivalent fuzzy measure, called L_δ-measure was proposed by our previous work as follows [11].

Definition 7: L_δ-measure

For given singleton measure $s\left(x\right)$, the composed measure of L-measure and δ-measure, denoted L_δ-measure, g_{L_δ}, is a multivalent fuzzy measure with determine coefficient $L\in\left[-1,\infty\right)$ on a finite set X, satisfying:

1) $\sum_{x\in X}s(x)=1$ (12)

2) $g_\delta\left(\phi\right)=0, g_\delta\left(X\right)=1, g_\delta\left(\left\{x\right\}\right)=s\left(x\right),\ \forall x\in X$ (13)

3) $\forall A\subset X, 1<\left|A\right|<\left|X\right|\Rightarrow$

$$g_{L_\delta}\left(A\right)=\begin{cases}\max_{x\in A}s\left(x\right) & if\ L=-1\\ \dfrac{\left[1+L\max_{x\in A}s\left(x\right)\right]\left(1+L\right)\sum_{x\in A}s(x)}{1+L\sum_{x\in A}s(x)}-L\max_{x\in A}s\left(x\right) & if\ L\in\left(-1,0\right)\\ \sum_{x\in A}s(x)+\dfrac{\left(\left|A\right|-1\right)L\sum_{x\in A}s\left(x\right)\left[1-\sum_{x\in A}s(x)\right]}{\left[\left|X\right|-\left|A\right|+\left(\left|A\right|-1\right)L\right]\sum_{x\in X}s\left(x\right)} & if\ L\in\left[0,\infty\right)\end{cases} \tag{14}$$

IV. TYPE II L_δ-MEASURE

L_δ-measure is the composed fuzzy measure of L-measure and the first half of δ-measure, here we consider the new composed fuzzy measure of L-measure and the whole of δ-measure, called Type II L_δ-measure as follows, it is more sensitive than L_δ-measure.

A. $L_{\delta'}$-measure

Definition 8: $L_{\delta'}$-measure

For given singleton measure $s(x)$, Type II composed measure of L-measure and δ-measure, denoted $L_{\delta'}$-measure, is a multivalent fuzzy measure with determine coefficient $L \in [-1,\infty)$ on a finite set X, satisfying:

1) $\sum_{x\in X} s(x) = 1$ (15)

2) $g_{\delta'}(\phi) = 0, g_{\delta'}(X) = 1, g_{\delta'}(\{x\}) = s(x), \forall x \in X$ (16)

3) $\forall A \subset X, 1 < |A| < |X| \Rightarrow$

$$g_{L_{\delta'}}(A) = \begin{cases} \max_{x\in A} s(x) & \text{if } L = -1 \\ \dfrac{\left[1 + L\max_{x\in A} s(x)\right](1+L)\sum_{x\in A} s(x)}{1 + L\sum_{x\in A} s(x)} - L\max_{x\in A} s(x) & \text{if } L \in (-1,1] \\ g_1(A) + \dfrac{(|A|-1)(L-1)\sum_{x\in A} s(x)\left[1 - g_1(A)\right]}{\left[|X| - |A| + (|A|-1)(L-1)\right]\sum_{x\in X} s(x)} & \text{if } L \in [1,\infty) \end{cases} \quad (17)$$

where $g_1(A) = \dfrac{2\sum_{x\in A} s(x)}{1 + \sum_{x\in A} s(x)}\left[1 + \max_{x\in A} s(x)\right] - \max_{x\in A} s(x)$ (18)

B. Important Properties of $L_{\delta'}$-measure

Theorem: Important Properties of $L_{\delta'}$-measure

1) $L_{\delta'}$-measure is a multivalent fuzzy measure with determine coefficient $L \in [-1,\infty)$
2), $L_{\delta'}$-measure is an increasing and continuous function of L on $[-1,\infty)$
3) if $L = -1$ then $L_{\delta'}$-measure is just the P-measure
4) if $L = 0$ then $L_{\delta'}$-measure is just the additive measure
5) if $-1 < L < 0$ then $L_{\delta'}$-measure is a sub-additive measure
6) if $0 < L < \infty$ then $L_{\delta'}$-measure is a supper-additive measure
7) If $\sum_{x\in X} s(x) = 1$ and $L = 0$ then L_δ-measure is just the λ-measure
8) P-measure, additive measure and λ-measure are the special cases of L_δ-measure

Proof. It is omitted for limitation of the space

V. Choquet Integral Regression Models

A. Choquet Integral

Definition 9: Choquet Integral [2-6]

Let μ be a fuzzy measure on a finite set X. The Choquet integral of $f_i : X \to R_+$ with respect to μ for individual i is denoted by

$$\int_C f_i d\mu = \sum_{j=1}^{n}\left[f_i\left(x_{(j)}\right) - f_i\left(x_{(j-1)}\right)\right]\mu\left(A^i_{(j)}\right), i = 1,2,\cdots,N \quad (19)$$

where $f_i\left(x_{(0)}\right) = 0$, $f_i\left(x_{(j)}\right)$ indicates that the indices have been permuted so that

$$0 \le f_i\left(x_{(1)}\right) \le f_i\left(x_{(2)}\right) \le \cdots \le f_i\left(x_{(n)}\right) \quad (20)$$

$$A_{(j)} = \left\{x_{(j)}, x_{(j+1)}, \cdots, x_{(n)}\right\} \quad (21)$$

B. Choquet Integral Regression Models

Definition 10: Choquet Integral Regression Models [7-15]

Let $y_1, y_2, \cdots, y_N$ be global evaluations of N objects and $f_1(x_j), f_2(x_j), \cdots, f_N(x_j), j = 1,2,\cdots,n$, be their evaluations of x_j, where $f_i : X \to R_+, i = 1,2,\cdots,N$.

Let μ be a fuzzy measure, $\alpha, \beta \in R$,

$$y_i = \alpha + \beta\int_C f_i dg_\mu + e_i,\ e_i \sim N(0,\sigma^2),\ i = 1,2,\cdots,N \quad (22)$$

$$(\hat{\alpha},\hat{\beta}) = \arg\min_{\alpha,\beta}\left[\sum_{i=1}^{N}\left(y_i - \alpha - \beta\int_C f_i dg_\mu\right)^2\right] \quad (23)$$

then $\hat{y}_i = \hat{\alpha} + \hat{\beta}\int f_i dg_\mu,\ i = 1,2,\cdots,N$ is called the Choquet integral regression equation of μ, where

$$\hat{\beta} = S_{yf} / S_{ff} \quad (24)$$

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} y_i - \hat{\beta}\frac{1}{N}\sum_{i=1}^{N}\int f_i dg_\mu \quad (25)$$

$$S_{yf} = \frac{\sum_{i=1}^{N}\left[y_i - \frac{1}{N}\sum_{i=1}^{N} y_i\right]\left[\int f_i dg_{\mu^*} - \frac{1}{N}\sum_{k=1}^{N}\int f_k dg_{\mu^*}\right]}{N-1} \quad (26)$$

$$S_{ff} = \frac{\sum_{i=1}^{N}\left[\int f_i dg_{\mu^*} - \frac{1}{N}\sum_{k=1}^{N}\int f_k dg_{\mu^*}\right]^2}{N-1} \quad (27)$$

VI. Experiment and Result

The total scores of 60 students from a junior high school in Taiwan are used for this research [10-11]. The examinations of four courses, physics and chemistry, biology, geoscience and mathematics, are used as independent variables, the score of the Basic Competence Test of junior high school is used as a dependent variable.

The data of all variables listed in Table III in our previous work [10-11] applied to evaluate the performances of five Choquet integral regression models with P-measure, λ-measure and δ-measure, L-measure measure, L_δ-measure and $L_{\delta'}$-measure based on γ-support respectively, a ridge regression model, and a multiple linear regression model by using 5-fold cross validation method to compute the mean square error (MSE) of the dependent variable. The formula of MSE is

TABLE I.
MSE OF REGRESSION MODELS

Regression model		5-fold CV MSE
Choquet Integral Regression model	measures	
	$L_{\delta'}$	47.8886
	L_{δ}	47.9742
	L	48.4610
	δ	48.7672
	λ	49.1832
	p	53.9582
Ridge regression		59.1329
Multiple linear regression		65.0664

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{28}$$

The experimental results of eight forecasting models are listed in Table I. It shows that the Choquet integral regression model with L_{δ} -measure based on γ-support outperforms other forecasting regression models.

CONCLUSION

In this paper, a multivalent composed fuzzy measure of L-measure and whole δ-measure, called Type II L_{δ} -measure, denoted $L_{\delta'}$ -measure, is proposed. This new measure is proved that it is of closed form with infinitely many solutions, and it can be considered as an extension of the two well known fuzzy measures, λ-measure and P-measure. Furthermore, this improved multivalent fuzzy measure is not only including the additive measure, but also having the same infinitely many measure solutions as L-measure. By using 5-fold cross-validation MSE, two experiments are conducted for comparing the performances of a multiple linear regression model, a ridge regression model, and the Choquet integral regression model with respect to P-measure, λ-measure, δ-measure, L-measure, L_{δ} -measure and the new fuzzy measure, $L_{\delta'}$ -measure based on γ-support respectively. The result shows that the Choquet integral regression models with respect to the proposed $L_{\delta'}$ -measure based on γ-support outperforms other forecasting models.

ACKNOWLEDGMENT

This paper is partially supported by the grant of National Science Council of Taiwan Government (NSC 98-2410-H-468-014).

REFERENCES

[1] A. E. Hoerl, R. W. Kenard, and K. F. Baldwin, Ridge Regression: Some Simulation, *Communications in Statistics*, vol. 4, No. 2, pp. 105-123, 1975.

[2] Z. Wang, and G. J. Klir, *Fuzzy Measure Theory*, Plenum Press, New York, 1992.

[3] Z. Wang, and G. J. Klir, *Generalized Measure Theory*, Springer Press, New York, 2009.

[4] M. Sugeno, *Theory of Fuzzy Integrals and its Applications*, unpublished doctoral dissertation, Tokyo Institute of Technology, Tokyo, Japan, 1974.

[5] L. A. Zadeh, *Fuzzy Sets and Systems*, vol. 1, pp. 3, 1978.

[6] G. Choquet, Theory of Capacities, *Annales de l'Institut Fourier*, vol. 5, pp. 131-295, 1953.

[7] H.-C. Liu, Y.-C Tu, C.-C. Chen, and W.-S. Weng, "The Choquet Integral with Respect to λ-Measure Based on γ-support", *2008 International Conferences on Machine Learning and Cybernetics*, Kunming, China, July 2008.

[8] H.-C. Liu, C.-C. Chen, Y.-D. Jheng, M.-F. Chien, "Choquet integral with respect to extensional L-measure and its application", *Proceedings of the 6th International Conference on Fuzzy Systems and Knowledge Discovery* (FSKD'09), pp. 131-136, 2009.

[9] H.-C. Liu, "A theoretical approach to the completed L-fuzzy measure", *2009 International Institute of Applied Statistics Studies* (2009IIASS) 2nd conference, Qingdao, China, July 24-29, 2009.

[10] H.-C. Liu, D.-B. Wu, Y-D Jheng, and T.-W. Sheu, "Theory of Multivalent Delta-Fuzzy Measures and its Application", *WSEAS Transactions on Info* H.-C. Liu, D.-B. Wu, Y-D Jheng, and T.-W. Sheu, "Theory of Multivalent Delta-Fuzzy Measures and its Application", *WSEAS Transactions on Information Science and Application,* Vol. 6, No. 6, pp. 1061-1070, July 2009.

[11] H.-C. Liu, C.-C. Chen, D.-B Wu, and T.-W. Sheu, "Theory and Application of the Composed Fuzzy Measure of L-Measure and Delta-Measures," *WSEAS Transactions on Information Science and control*, Issue 8. Vol. 4, pp. 359-368, August 2009.

[12] J.-I Shieh, H.-H. Wu, H.-C. Liu, Applying Complexity-based Choquet Integral to Evaluate Students' Performance, *Expert Systems with Applications*, 36, pp. 5100-5106, 2009.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 205-208

Research on Distributed Geo-Computing Oriented Self-organized P2P Network

Xicheng Tan [1], and Fang Huang [2]
[1]International School of Software, Wuhan University, Wuhan, China
Email: xichengtan@gmail.com
[2] Institute of Geo-Spatial Information Technology, College of Automation, University of Electronic Science and Technology of China, Chengdu, P.R. China
Email: hfhbhzp@uestc.edu.cn

Abstract—**With the extending of spatial information system into the distributed network environment, it faces some challenges including the mass data character of the spatial data, the limited band width of current network, the devilishly centralized spatial information management and geographic computing resources, as well as the higher requirements of the spatial information service capability. For overcoming these challenges, this paper puts forward a Geo-Computing oriented self-organized P2P network model, and the structure of the P2P network is designed. For performing spatial analysis tasks, the paper also analyzes the spatial data management on the self-organized P2P network. Finally, the test system, which has simulated the slope analyzing based on the self-organized P2P network, is also presented. Compare with the single server based spatial analysis systems the P2P computing based analysis task performs more efficiently and has a better capability of supporting huge amount of requests from the users.**

Index Terms—**Spatial Analysis, Distributed Computing, P2P Computing, High Performance Computing (HPC), Task Dispatch**

I. INTRODUCTION

Geo-Computing is art and science of solving the complex spatial issue with the computer. Computing science use the computer to research science issue, and the High Performance Geo-Computing (HPGC) is a use field of the High Performance Computing. With the high performance resources HPGC can play great role in solving the problems of geosciences. The use fields of HPGC include spatial data analysis, Dynamic modeling, Simulation, Temporal-Spatial science, Visualization, 3DGIS and VR, etc. However, most researches of HPGC are focused on the parallel computing algorithms on the computer clusters currently[1,2,3,4]. With developing of Geo-Computing extents to the distributed environment, the parallel computing on Grid environment has been a new research direction[5,6].

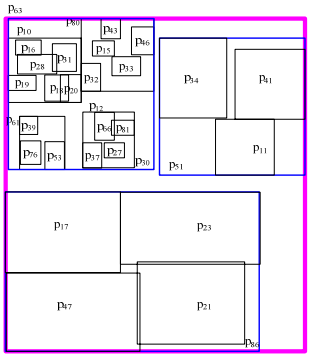
With the developing of Peer-to-Peer (P2P) network, it has given us a more powerful technique to resolve the issue of HPGC. P2P computing has been an important model of distributed computing, and it can operate the collaborative computing by using the spare computing resource of the internet. Moreover, compare with the GRID computing, P2P is a kind of thin-core computing model, and it is benefit for using the distributed computing resources[8, 9]. There are many researches based on P2P computing, such as SETISETI@home[10]，XtremWeb, Chord, etc. The idea of using spare computing resources has been addressed for some time by traditional distributed computing systems. The *Beowulf* project from NASA was a major milestone that showed that high performance can be obtained by using a number of standard machines.

In this paper we analysis the properties of spatial data distributed store. Based on these properties, we put forward a P2P overlay network structure based on R-Tree indexed spatial data, and the topological structure of the P2P network is presented also. More over, the Logical Model and Task Scheduling of Self-Organized P2P Geo-Computing are analyzed. Finally, the test system, which has simulated the slope analyzing based on the P2P computing model, is also presented.

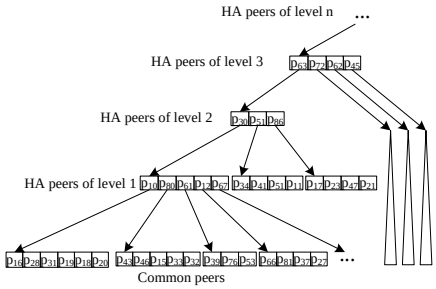
II. SELF-ORGANIZED P2P OVERLAY NETWORK STRUCTURE

In order to distribute and manage the distributed spatial data conveniently and effectively, it is crucial to select one feasible type of the P2P network firstly. Among those spatial data, there exist uncountable spatial relations, which play important roles in the data processing on specific P2P network, mainly including the spatial data locating and indexing. Due to those reasons, the methods for spatial data storage are varied from that of the non-spatial data storage. If the storage of spatial data adopts the type of the non-spatial data, the efficiency of spatial data location and transferring will fall remarkably. As a result, the performance of the whole system will also be debased. From this point, the method of spatial data storage, especially the spatial data locating and indexing, need to draw much attention in P2P network. Unfortunately, the existing P2P applications consider litter on how to locating and indexing the spatial data.

In the managing of spatial data in the P2P network, the most important work is how to index the distributed spatial data in the P2P network. Since abundant spatial relations are contained in the spatial data, it is more convenient, to some extents, to locate and manage the resource in the P2P network if the spatial relations are used adequately. We adopt spatial data R-tree index to improve the efficiency and correctness of whole resource location as show in Fig1(a), and a P2P overlay network structure is constructed based on the spatial data R-Tree index. The R-Tree indexed spatial data and the P2P overlay network structure are shown as Fig.1(b)



(a) R-tree indexed spatial data



(b) Spatial data r-tree based P2P network structure

Figure 1. P2P overlay network structure based on r-tree indexed spatial data.

R-Tree is a balance tree and it has two kinds of nodes, leave node and non-leave node. Each node has some index items. To the former, the index item points to the Minimum Bounding Rectangle (MBR) of the spatial data kept by the leave node; To the latter, the items also points to one specific MBR that contains the MBRs of the leave nodes. As shown in Fig.1 (a), *p16*, *p28*, *p31*, *p19*, *p18*, *p20* are leave nodes, while *p10* is non-leave node, in which the item contains the MBRs of *p16 et al.*.

Fig.1 (b) shows the spatial data R-Tree based self-organized P2P structure. In the structure, an important component--hierarchical Agent (HA) is used, and all the HA peers are the non-leave nodes in the R-Tree index. Thus HA peers of every level can manage some common peers through MBR of spatial data kept by them. In the structure, HA peer of level *i* (HAi) is the core of the peer cluster (or group) to which HAi-1 or the common peers belong. If the MBR of one HA1 contains the MBRs of common peers, these common peers will join in the same group and connect to the HA1, and will be managed by the HA1. If some MBRs in HA1 peers contains the MBR of a HA2 peer, these HA1 peers will join a same group and connect to the HA2, and will be managed by the HA2. In this way, all peers can be managed by HA on different levels.

In order to enhance the routing efficiency of the P2P network, Peers in the same cluster will construct a ring shown in Fig.2. The service state information of the peers can be ensured by regular detecting along the ring. For this purpose, a peer will send detecting message regularly to the previous peer to ensure the peer is on-line and to get the network condition of the peer. The regular detecting reduces the efficiency depletion of routing maintaining.

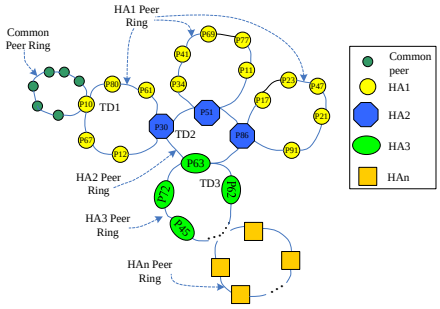


Figure 2. Topological structure of the P2P network.

III. THE DISTRIBUTED SPATIAL DATA MANAGEMENT ON THE P2P NETWORK

All of the HA peers and common peers are spatial data storage peers as well as client peers. HA peers require higher equipment performance, internet bandwidth and stable online service. Both HA peers and the common peers contribute a part of storage space, and then the contributed storage space will be managed and checked via routing function of P2P network. Comparing with Mixed-Structure P2P network, there isn't server peer in this network, which can increase the stability of the system, as well as keeping its efficiency.

HA Peers has stable service time, fixed public network IP, biggish bandwidth, higher performance. If some spatial data peers can reach certain conditions, it can be joined the network as a HA peer too. Ring structure constructed by common peers is to secure the great amount of grouped common peers. Every HA peer maintains a common peer's ring, common peers in this ring has higher spatial relation, i.e. these common peers are accepting spatial data service of the same or neighboring space. While accepting service, a great amount of common peers are also the contributors offering the spatial information. Common peers in a same ring could set links with one another and transfers the needed spatial data.

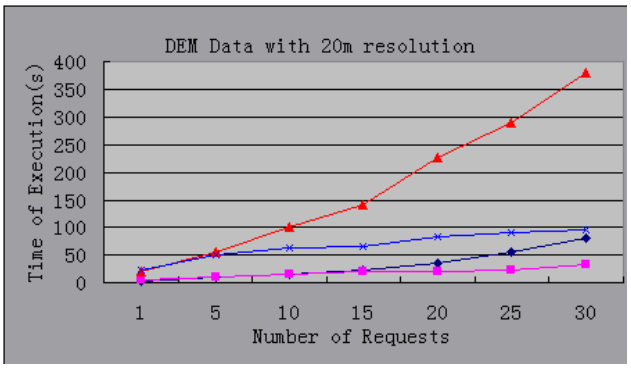
Another issue is deciding the information kept by different peers. For the P2P network is constructed based on the R-Tree of the spatial data kept by the peers, undoubtedly every peer should keep the spatial data as well as the MBR of the spatial data. Meanwhile, each peer manages the metadata of the kept spatial data, and with the metadata, the information of the kept spatial data such as data classes, precision information etc. can be described. For regular detecting in the ring, each peer keep the ID of the previous and the subsequent peer in the ring. Moreover, HA peer keeps IDs of the common peers managed by itself and its subordinate HA peers.

IV. PERFORMANCE EVALUATION

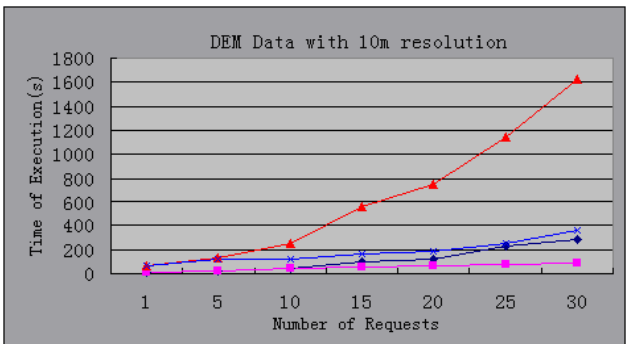
For evaluating the performance of task assignment and the geo-computing application in the P2P network, we take a evaluating scenario to evaluate performance of geo-Computing applications.

For evaluating the performance of task assignment and the geo-computing application in the P2P network, we take two evaluating scenarios. The first scenario is used to evaluate the Performance of Geo-Computing Task Assignment and the second is to evaluate performance of geo-Computing applications.

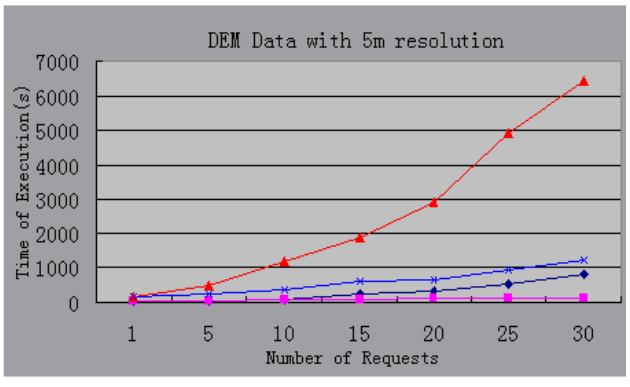
In this scenario, we compare the performance of the average slope value computing on the P2P based application and the C/S based application. There are 10 HA peers and 39 common peers in P2P network. The HA peers and the common peers all are the PC with 2.7G CPU, 512MB ROM and 160G disk. A single HA peers ring with 2 levels of HA peers is constructed and therefore there 10 HA peers in the ring, the number of common peers managed by each HA peer varies from 3 to 5. Three DEM (Digital Elevation Model) data files with different resolution of, respectively, 20m, 10m and 5m, is stored in the P2P network distributed according to the R-Tree index.



(a) Execution time of analysis on DEM with 20m resolution



(b) Execution time of analysis on DEM with 10m resolution



(c) Execution time of analysis on DEM with 5m resolution

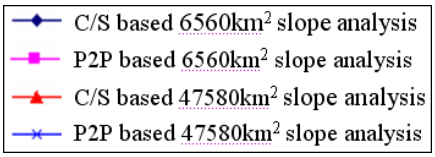


Figure3 The Comparison of the performance of the slope analysis based on different computing model

The environment of C/S based geo-computing application mentioned as follows. The server machine with 3G CPU and 2G ROM, and all the client machines' parameters are the same with that of the client peers in the P2P environment. All the three DEM data files are stored on the server machine.

There are three conditions in the test: (1) task extent; (2) number of concurrent tasks; (3) resolutions of the DEM data files. All the three conditions can be set by the client application.

We set two different task extents: 6560km^2 and 47580km^2. We define the task with 6560km^2 as process A, and task with 47580 km^2 as process B. For both process A and B, the three levels of resolutions are used. The number of concurrent tasks varies from 1 to 30. The performance of the average slope value computing is shown in Fig.3. With the increasing of the DEM resolution, the extending of the task extent and the rising of the concurrent requests number, the execution time of C/S based computing increases more sharp than that of P2P based computing even though the data is distributed stored in the P2P network. Because more HA peers and common peers act as workers in the geo-computing during the execution of tasks, besides, the communication load of the server in the C/S based computing is bigger than that of the P2P network.

V. CONCLUSION

This paper presented a spatial data R-Tree index based P2P structure firstly, and the mechanism of the spatial data management is designed. According to the P2P geo-computing logical view, the geo-computing oriented task scheduling of P2P geo-computing is designed.

Simulation results show that the P2P geo-computing job assignment gets considerable performance. The comparison of the geo-computing applications performance shows that the P2P based geo-computing is more excellent than the C/S based geo-computing.

ACKNOWLEDGMENT

The research work is supported by the Geographic Spatial Information Engineering Laboratory of China State Bureau of Surveying and Mapping (No. 200806)

REFERENCES

[1] Mineter, M.J., 2003. A software framework to create vector-topology in parallel GIS operations, International Journal of Geographical Information Science, 17(3): 203-222.

[2] Wang, F.J., 1993. A parallel intersection algorithm for vector polygon overlay. Computer Graphics and Applications 13(2): 74-81.

[3] Xiong, D., and Marble, D.F., 1996. Strategies for Real-Time Spatial Analysis Using Massively Parallel SIMD

Computers: An Application to Urban Traffic Flow Analysis. International Journal of Geographical Information Systems, 10(6): 769-89.

[4] Armstrong, M. P., and Densham, P. J., 1992. Domain Decomposition for Parallel Processing of Spatial Problems. Computers, Environment, and Urban Systems, 16: 497-513.

[5] Huang, F., 2009. Implementation and QoS for high-performance GIServices in spatial information grid.

[6] Wang, L.Z., Chen, J.J., et al (ed.), Quantitative Quality of Service for Grid Computing: Applications for Heterogeneity, Large-Scale Distribution and Dynamic Environments. New York: IGI Global, pp.181-203.

[7] B. E. W. Garces E L , Felber P A Hierarchical Peer-to-Peer systems,Parallel Processing Letters 2003: 13(4), 643-657.

[8] G. Jon B. Weissman, Network Partitioning of Data Parallel Computations[C], presented at Proceeding of Third IEEE International Symposium on High Performance Distributed Computing, 1994.

[9] T.-H. K. J. M.Purtilo, Load Balancing for Parallel loops in workstation clusters, Technical Report,Department of Computer Science,University of Maryland[J], 1996.

[10] J. C. David P Anderson , Eric Korpela, SETI @home : An experiment in public-resource computing

[11] http://setiathome.ssl.berkeley.edu/cacm/cacm.html, 2002.

[12] Xicheng Tan, Liang Yu, Fuling Bian. Large-scale P2P network based distributed virtual geographic environment (DVGE).prensented at proc of Geospatial Information Technology and Applications.2007:6754(2),675427-675430

[13] H. Jin, F. Luo, X. F. Liao, Q. Zhang, and H. Zhang, Constructing a P2P-based high performance computing platform, in Computational Science - Iccs 2006, Pt 4, Proceedings, vol. 3994, Lecture Notes in Computer Science, 2006, 380-387.

[14] J. Z. R. D. McLeod, Application layer routing options for efficient data transfer over the Internet[C], presented at Proc of 2002 IEEE Canadian Conf on Electrical & Computer Engineering, Los Alamitos, 2002.

[15] I. F. Karl Czajkowski , Nicholas Karonis, A resource management architecture for metacomputing systems http://www.globus.org, 2003.

[16] H. E.-R. a. T. G. Lewis, Scheduling Parallel Program Tasks onto Arbitrary Target

[17] Machines[J], Journal of Parallel and Distributed computing, 1990: 9(1), 138-153.

[18] T. Y. A. Gerasoulis, Scheduling Parallel Tasks on an Unbounded Number of Processors[J], IEEE Transaction on Parallel and Distributed System, 1994: 5(9) ,951-967.

[19] Yu, Song,Xue, B. A. I.,Shuchun, J. U.,Xiujuan, H. A. N. 2005, Building Dynamic GIS Services based on peer-to-peer, Semantics, Knowledge and Grid, 2005. SKG '05, 68-68.

[20] N. Preguiça, M. Shapiro, C. Matheson. Semantics-based reconciliation for collaborative and mobile environments. *CoopIS Conf.*, 2003.

[21] S. Ratnasamy et al. A scalable content-addressable network. Proc. of SIGCOMM, 2001.

[22] SETI@home. http://www.setiathome.ssl.berkeley.edu/.

[23] M. Shapiro. A simple framework for understanding consistency with partial replication.Technical Report, Microsoft Research, 2004.

[24] Stoica et al. Chord: A scalable peer-to-peer lookup service for internet applications. *Proc.of SIGCOMM*, 2001.

[25] Tanaka, P. Valduriez. The Ecobase environmental information system: applications,architecture and open issues. *ACM SIGMOD Record*, 3(5-6), 2000.

[26] Tatarinov et al. The Piazza peer data management project. *SIGMOD Record 32(3)*, 2003.

[27] Tomasic, L. Raschid, P. Valduriez. Scaling access to heterogeneous data sources with DISCO. *IEEE Trans. on Knowledge and Data Engineering*, 10(5), 1998.

[28] P. Valduriez: Parallel Database Systems: open problems and new issues. *Int. Journal on Distributed and Parallel Databases*, 1(2), 1993.

[29] Yang, H. Garcia-Molina. Designing a super-peer network. *Int. Conf. on Data Engineering*,2003.

[30] W. Nejdl, W. Siberski, M. Sintek. Design issues and challenges for RDF- and schemabased peer-to-peer systems. *SIGMOD Record, 32(3),* 2003.

[31] B. Ooi, Y. Shu, K-L. Tan. Relational data sharing in peer-based data management systems. *SIGMOD Record, 32(3),* 2003.

[32] T. Özsu, P. Valduriez. *Principles of Distributed Database Systems*. 2nd Edition, PrenticeHall, 1999.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 209-212

On Software Development for Electric Power Steering System Based On uCOS-II

Bing Zhou[1], and Fengmei Hou[2]
[1] School of computer science and engineering, Jiangsu Teachers University of Technology, Changzhou 213001, China
Email: zhb@ jstu.edu.cn
[2] Modern Education Technology Center, Yangzhou Polytechnic College, Yangzhou 225000, China
Email: liuhm1128@yahoo.com.cn

***Abstract*—Electric Power Steering (EPS) is a full electric system which reduces the amount of steering effort by directly applying the output from an electric motor to the steering system and has attracted much attention for their advantages. Controller design and software development of EPS is one of the key technologies that are critical to the system's design. The constitutions and its operational mechanism of electric power steering system were introduced, and the EPS hardware framework based on the ARM microprocessor was presented and the software development based on usos-ii was designed. The tests were performed and the results confirmed that the system was stable and credible, and can meet the requirements of steering performance. It has practical engineering significance to the implement of EPS motor control strategy, to the improvement and optimization of EPS function and to the steering manipulation safety.**

***Index Terms*—Electric Power Steering; EPS; ARM; usos-ii; controller; software**

I. INTRODUCTION

With the development of system science and system dynamics and requirements for environmental protection and energy conservation, and for the driving comfort and security, Electric power steering (EPS) system has attracted much attention for their advantages with respect to improved fuel economy and have been widely researched and adopted as automotive power steering equipment in recent years. Controller design and software development of EPS is one of the key technologies that are critical to the system's design. In this paper, the constitutions and its operational mechanism of electric power steering system were introduced, and the EPS hardware framework based on the ARM microprocessor was presented and the software development based on usos-ii was designed. The tests were performed and the results confirmed that the system was stable and credible, and can meet the requirements of steering performance.

II. EPS CONSTITUTIONS AND OPERATIONAL MECHANISM

A column-type EPS is shown in Figure 1. It consists of a torque sensor, which senses the driver's movements of the steering wheel as well as the movement of the vehicle; an ECU, which performs calculations on assisting force based on signals from the torque sensor and vehicle sensor; a motor, which produces turning force according to output from the ECU; and a reduction gear, which increases the turning force from the motor and transfers it to the steering mechanism [1, 2].

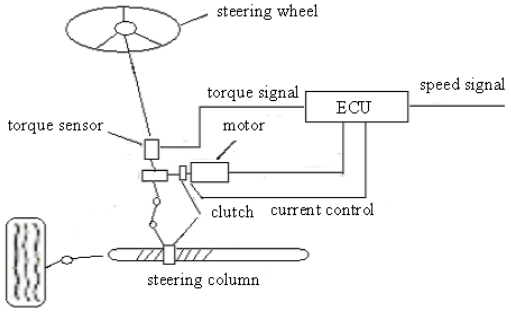


Figure 1. A column-type EPS system

The main purpose of electric power steering system is, of course, to provide assist to the driver. This is achieved by the torque sensor, which measures the driver's torque and sends a signal to the controller proportional to this torque. The torque information is processed in the controller and an assist command is generated. This assist command is further modulated by the vehicle speed signal, which is also received by the controller. This command is given to the motor, which provides the torque to the assist mechanism. The gear mechanism amplifies this torque, and ultimately the loop is closed by applying the assist torque to the steering column.

III. INTRODUCTION

A. Motor control principle

EPS motor is a permanent magnetic field DC motor. Attached to the power steering gear assembly, it generates steering assisting force. Figure 2 is the schematic of the electric circuit, including the windings resistance R and inductance L [3, 4].

We can get the transfer function of a DC motor, and the mathematical model is given by

$$\begin{cases} V_{\mathrm{m}} = R_{\mathrm{m}} i_{\mathrm{m}} + L_{\mathrm{m}} \dfrac{\mathrm{d}i_{\mathrm{m}}}{\mathrm{d}t} + e_{\mathrm{m}} \\ T_{\mathrm{m}} = K_{\mathrm{t}} i_{\mathrm{m}} \\ T_{\mathrm{m}} = T_{\mathrm{L}} + T_{\mathrm{f}} \\ e_{\mathrm{m}} = K_{\mathrm{e}} \omega_{\mathrm{m}} \end{cases} \quad (1)$$

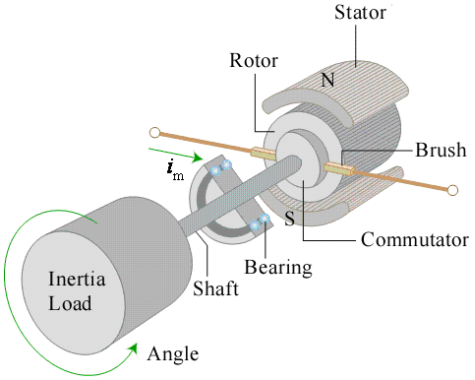


Figure 2. DC Motor Construction

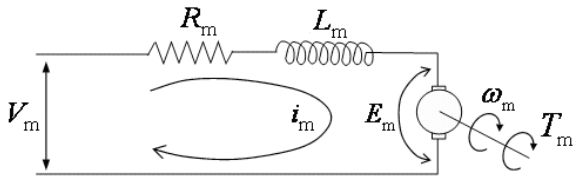


Figure 3. Electric Circuit

The mathematical model in the frequency domain are

$$\begin{cases} V_m(s) = R_m I_m(s) + L_m s I_m(s) + E_m(s) \\ J_m s\omega_m(s) = -B_m\omega_m(s) - T_L(s) + T_m(s) \\ E_m(s) = K_m\omega_m(s) \\ T_m(s) = K_m I_m(s) \end{cases} \quad (2)$$

And, there is

$$T_m(s) = \frac{K_m(J_m s + B_m)}{(L_m s + R_m)(J_m s + B_m) + K_m{}^2} V_m(s) + \frac{K_m{}^2}{(L_m s + R_m)(J_m s + B_m) + K_m{}^2} T_L(s) \quad (3)$$

$$= \begin{bmatrix} M_1 & M_2 \end{bmatrix} \begin{bmatrix} T_L(s) \\ V_m(s) \end{bmatrix}$$

Where,

$$M_1 = \frac{K_m{}^2}{(L_m s + R_m)(J_m s + B_m) + K_m{}^2} \quad (4)$$

$$M_2 = \frac{K_m(J_m s + B_m)}{(L_m s + R_m)(J_m s + B_m) + K_m{}^2} \quad (5)$$

B. EPS control hardware framework

The block diagram of control strategy of EPS system dynamics is shown in Figure 4 and 5. The assist characteristic unit determines the reference current to the motor based on the driving conditions, and the controller computes the control signal which minimizes the error between and the actual current. And the port resource allocation table is shown in Table 1.

The EPS controller consists of an interface circuit that coordinates the signals from the various sensors, an A/D converter and a PWM unit that are all built into an one-chip microprocessor, a watchdog timer circuit that monitors the operation of this microprocessor, the motor-drive circuit that consists of power MOSFETs in an H bridge circuit driven by pulse width modulation over a 20kHz carrier. The ECU conducts a search for data according to a table lookup method based on the signals input from each sensor and carries out a prescribed calculation using this data to obtain the assist force [5].

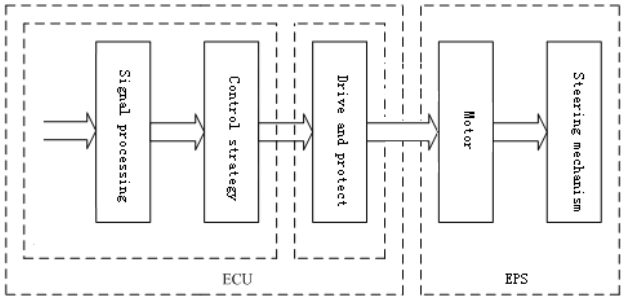


Figure 4. EPS control strategy

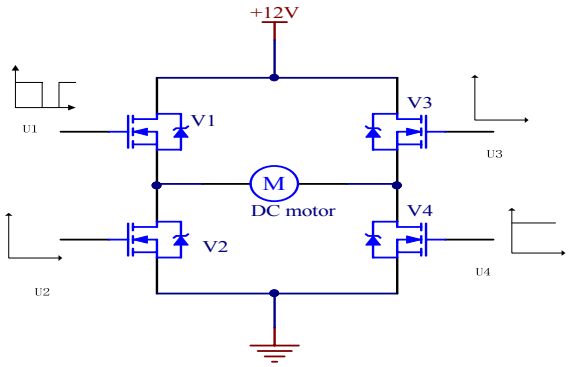


Figure 5. Motor drive circuit

The microprocessor LPC2131 is based on a 16/32 bit ARM7TDMI-S™ CPU with real-time emulation and embedded trace support, together with 128/256 kilobytes of embedded high speed flash memory. A 128-bit wide memory interface and unique accelerator architecture enable 32-bit code execution at maximum clock rate. With their compact 64 pin package, low power consumption, various 32-bit timers, 4-channel 10-bit ADC, 2 advanced CAN channels, PWM channels and 46 GPIO lines with up to 9 external interrupt pins these microcontrollers are particularly suitable for automotive and industrial control applications as well as medical systems and fault-tolerant maintenance buses. With a wide range of additional serial communications interfaces, they are also suited for communication gateways and protocol converters as well as many other general-purpose applications.

Pulse width modulation (PWM) is a powerful technique for controlling analog circuits with a processor's digital outputs. PWM is employed in a wide

variety of applications, ranging from measurement and communications to power control and conversion [6].

The PWM waveform, p(t) is defined as

$$p(t) = \begin{cases} 1 & r(t) \geq 0 \\ -1 & else \end{cases}. \quad (6)$$

Where, $w(t) = \alpha |r(t)|$ and $kT \leq t \leq (k+1)T$.

TABLE I.
PORT RESOURCE ALLOCATION TABLE

Input signal	AD0.1	P0.28	Torque 1
	AD0.3	P0.30	Torque 2
	AD0.5	P0.26	Motor current
	GPIO	P0.17	Vehicle speed
	GPIO	P0.18	Engine speed
Output signal	Clutch	P0.27	Torque
	PWM	P0.8	H-bridge 1
		P0.23	H-bridge 2
		P0.9	H-bridge 3
		P0.25	H-bridge 4

Ⅳ. SOFTWARE DEVELOPMENT ARCHITECTURE

Software development architecture is based on uCOS-II (Micro Controller Operation System version 2) which is shown in Figure 6. uCOS-II is an open source, real-time operating system kernel intended for real-time embedded applications. The three parts associated with the processor are OS_CPU_C.C, OS_CPU.H, OS_CPU_A.ASM. uCOS-II is a multi-tasking real time operating system, task is a process including a certain priority level, a separate set of CPU registers and stack space, its status includes sleep state, ready state, running state, suspend and interrupted state. The task can be an infinite loop, it can be removed after completion of the first performance, with a return type and a parameter, but the return type must be defined as void type.

For EPS software system, it is an event-driven multi-task management system and be classified into multiple tasks. The tasks will be different for each operating state and switched to the next state to achieve the operation of system dynamics. The structure of multi-task is as following [7, 8].

```
void UserTask (void *pdata)
{
  for (;;)
    {
      OSMboxPend();
      OSQPend();
      OSSemPend();
      OSTaskDel(OS_PRIO_SELF);
      OSTaskSuspend(OS_PRIO_SELF);
      OSTimeDly();
      OSTimeDlyHMSM();
    }
}
```

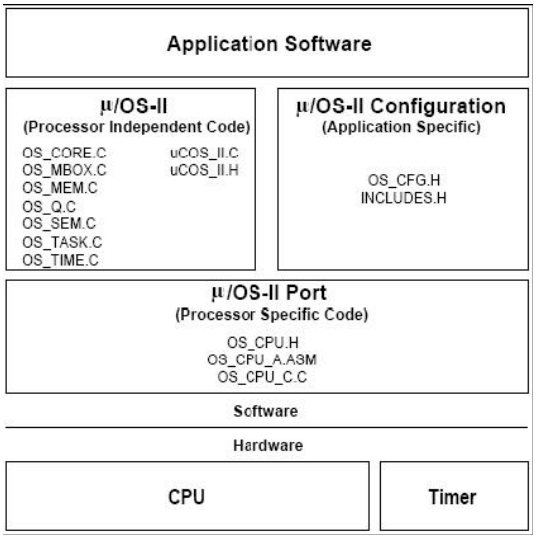


Figure 6. Software development architecture

The task is created as the following:

```
Taskcreate (INT8U prio, OS_STK *ptos, OS_STK *pbos, INT16U id, INT16U stk_size, void *pext, INT16U opt)
{
  OS_TCB *ptcb;
  OS_ENTER_CRITICAL();
  ptcb = OSTCBFreeList;
  if (ptcb != (OS_TCB *)0)
    {
    OSTCBFreeList= ptcb->OSTCBNext;
    OS_EXIT_CRITICAL();
    ptcb->OSTCBStkPtr= ptos;
    ptcb->OSTCBPrio= (INT8U)prio;
    ptcb->OSTCBStat= OS_STAT_RDY;
    ptcb->OSTCBDly= 0;
    pext= pext;
    stk_size= stk_size;
    pbos= pbos;
    opt= opt;
    id= id;
    ptcb->OSTCBY= prio >> 3;
    ptcb->OSTCBBitY= OSMapTbl[ptcb->OSTCBY];
    ptcb->OSTCBX= prio & 0x07;
    ptcb->OSTCBBitX= SMapTbl[ptcb->OSTCBX];
    OS_ENTER_CRITICAL();
    OSTCBPrioTbl[prio] = ptcb;
      ptcb->OSTCBNext= OSTCBList;
      ptcb->OSTCBPrev= (OS_TCB *)0;
      if (OSTCBList != (OS_TCB *)0)
        {
         OSTCBList->OSTCBPrev = ptcb;
        }
      OSTCBList= ptcb;
      OSRdyGrp|= ptcb->OSTCBBitY;
      OSRdyTbl[ptcb->OSTCBY] |= ptcb->OSTCBBitX;
      OS_EXIT_CRITICAL();
      return (OS_NO_ERR);
      }
    else
```

```
        {
      OS_EXIT_CRITICAL();
      return (OS_NO_MORE_TCB);
        }
  }
```

For each task, there are ADC signal collection task, PWM control task, interrupt control task, they are the following:

(1) The PWM task is as the following:

```
void PWMTask ( )
{
  PWMPR= 0x00;
  PWMMCR= 0x02;
  PWMPCR= 0x0400;
  PWMMR0= Fpclk / 1000;
  PWMMR2= PWMMR0 / 2;
  PWMLER= 0x05;
  PWMTCR= 0x02;
  PWMTCR= 0x09;
  PWM0_F=Fpclk/2000;
  PWMPR= 0x00;
  PWMMCR= 0x02;
  PWMPCR= 0x5000;
  PWMMR0= PWM0_F;
  PWMMR4= PWM4_Q1;
  PWMMR6= PWM6_Q3;
  PWMLER= 0x51;
  PWMTCR= 0x02;
  PWMTCR = 0x09;
}
```

(2) The Interrupt task is as the following:

```
void InterruptTask ( )
{
  VICIntSelect= 0x00000000;
  VICVectCntl0= 0x20 | 14;
  VICVectAddr0= (uint32)IRQ_Eint0;
  VICIntEnable= 1 << 14;
  EXTINT = 0x01;
  VICVectAddr = 0;
}
```

(3) The ADC task is as the following:

```
void ADCTask ( )
{
  uint32 ADC_Data;
  AD0CR = (1 << m_or_s)
  ((Fpclk / 1000000 - 1) << 8)
     (0 << 16)
     (0 << 17)
     (1 << 21)
     (0 << 22)
     (1 << 24)
     (0 << 27);
     DelayNS(10);
  while ((AD0DR & 0x80000000) == 0);
     ADC_Data = AD0DR;
     ADC_Data = (ADC_Data >> 6) & 0x3ff;
     return(ADC_Data);
}
```

Ⅴ. SOFTWARE DEVELOPMENT ARCHITECTURE

In order to test the performance of the design of hardware and software system, the testing program is made. The PWM signal is generated from controlling analog circuits with a processor's digital outputs. Through the use of high-resolution counters, the duty cycle of a square wave is modulated to encode a specific analog signal level. The voltage source is supplied to the analog load by means of a repeating series of on and off pulses. The on-time is the time during which the DC supply is applied to the load, and the off-time is the period during which supply is switched off. Given a sufficient bandwidth, any analog value can be encoded with PWM. Figure 7 shows the PWM outputs maintain their output voltage stability and speed stability at different load conditions. The results show that control method can be realized easily and the system devised is stable and credible, and can meet the requirements of steering performance.

The demands for faster speed, higher quality, and reduced power requirements in vehicles are continually increasing. In order to respond to these demands, research and development is under way on the application of electronic control with the aim of further improving functions and performance.

REFERENCES

[1] Jiang Haobin et al.Hardware design and experiment research of automotive electric power steering system. The 3rd China-Japan Conference on Mechatronics 2006 Fuzhou,2006,68-71.

[2] Aly Badawy et al.Modeling and analysis of an electric power steering system.SAE paper 1999-01- 0399.

[3] Tanaka.Motors for electric power steering. Technical reports.2003.

[4] International Rectifier.HV Floating MOS-Gate Driver IC's.AN978.

[5] Ronald K. Jurgen. Automotive electronics handbook [M], Second edition, McGraw-Hill, Inc, 1999.

[6] Zhao Jingbo, Chen Long, Jiang Haobin, et al. Design and full-car tests of electric power steering system. Computer and Computing Technologies in Agriculture. United States: SPRINGER, 2008: 729-736.

[7] International Rectifier.Bootstrap Component Selection For Control IC's.DT98-2a.

[8] Richard Valentine. Motor Control Electronics Handbook [M], McGraw-Hill, 1998.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 213-217

A New Family of Cayley Graph Interconnection Networks Based on Wreath Product

Zhen Zhang [1,2], and Xiaoming Wang[3]
[1] Department of Computer Science, South China University of Technology
[2] Department of Computer Science, Jinan University, Guanghzou, China
zhang2003174@yahoo.com.cn
[3] Department of Computer Science, Jinan University , Guangzhou, china

***Abstract*—In this paper, we propose a new class of Cayley graph WG_n^{2m} for building interconnection networks with fixed degree of *m*-3(or *m*-2 when *n*=2). We present a routing algorithm for the Cayley graph WG_n^{2m} with the diameter upper bounded by $\lfloor 5n/2 \rfloor$. Some embedding properties is also derived.**

Index Terms—**interconnection networks, cayley graph, routing algorithm, diameter, network embedding.**

I. INTRODUCTION

Design of interconnection networks is an important integral part of any parallel processing or distributed systems. Performance of the distributed system is significantly determined by the choice of the network topology. Desirable properties of an interconnection network include low degree, low diameter, symmetry, low congestion, high connectivity, and high fault tolerance. For the past several years, there has been active research on a class of graphs called Cayley graphs because these graphs possess many of the above properties.

A Cayley graph can be explained as follows: Let $S=\{s_1,s_2,\ldots,s_m\}$ be a set of generators for a finite group G. Then, $\Gamma=(V,E)$ is Cayley graph, where each member of the set of vertices V corresponds to an element of the G and an undirected edge(v_i,v_j) is a member of E if and only if there exists a generator $s_k \in S$ such that $v_i s_k=v_j$, The generator set S have no identity element I and be closed under inverses so that the graph has no loops and can be viewed as being undirected. Many well-known networks are Cayley graphs, such as *Hypercube*[4], *CCC*[12] and *k-ary n-cubes* [4].

Every Cayley graph is vertex transitive [1][3][5], that means , for every pair of vertices *a* and *b*, there exists an automorphism of the graph that maps *a* into *b*. An attractive feature of vertex transitive graphs is that routing between two arbitrary vertices reduces to routing from an arbitrary vertex to the identity node.

A class of Cayley graphs based on permutation groups has proven to be suitable for designing interconnection networks, such as *Star graph*[1][2][13], *Hypercubes*[4], *Pancake gaphs*[2][16], *Shuffel-Exchange Permutation Network*[10], the *Rotation-Exchange Network*[21]. These graphs are symmetric, regular, and seem share the desirable properties described above. Vadapalli and Srimani [15] prposed the trivalent Cayley graph with a fixed degree of three. This class of graphs is known to have logarithmic diameter and maximal fault tolerance [11][14][15][17][18]. Then, they proposed a family of Cayley graph of constant degree four, which is isomorphic to the wrapped around butterfly graph[16]. Fu and Chau proposed cylic-cubes, which are Cayley graph with fixed degrees being any even number greater than or equal four[8]. These graphs have optimal fault tolerance and logarithmic diameters. The shortest path routing and embedding of a Hamiltonian cycle, meshes, and hypercubes are also discussed. R.K.Das and B.P.Sinha develop a new network topology with odd degrees [7]. The point-to-point routing and one-to-all routing algorithms are also developed . More recently, Zhou, Du and Chen propose a family of Cayley graph of odd fixed degrees[22]. Hsieh and Hsiao develop a new family Cayley graph $G_{k,n}$ with *k*-degree[9], which possess a useful property in that the degree of each node is fixed by a general positive integer *k* without regard to the number of nodes.

In this paper, we generalize the concept of fixed degree and propose a new family of Cayley graphs WG_n^{2m}. These graphs possess a very useful property in that the degree of each is fixed by a general positive integer without regard the number of nodes. This article is organized as follows. In section II, we introduces the topology of WG_n^{2m}. Section III is devoted to dealing with a routing algorithm and establishing the upper bound of its diameter. The graph-embedding properties is discussed in Section IV. Finally, we give some conclusions in Section V.

II. CAYLEY GRAPH OF WG_n^{2m}

In this section, we give the definition of Cayley graph WG_n^{2m}. First, we define the ranked symbol system as follows: let $t_1,t_2,\ldots,t_n$ be n different symbols, and 2*m* ranks 0,1,2,…, 2*m*-1, where $n\geq 2$ and $m\geq 1$. We assign each symbol a rank, such as t_j^i (assigning rank *i* to symbol t_j), where $0\leq i\leq 2m-1$. For *n* distinct symbols, there are *n* different cyclic permutations of the symbols. Since each symbol can be represented in 2*m* distinct ranked forms, the vertex set of the WG_n^{2m} has a cardinality of $n\cdot(2m)^n$. Let *I* denote the identity permutation of $t_1^0 t_2^0 \ldots\ldots t_n^0$, and

$a_1^{i_1}a_2^{i_2}.....a_n^{i_n}$ denote an arbitrary vertex. Each edge of WG_n^{2m} is of type $(v, \delta(v))$, where $\delta \in \{f, f^{-1}, g, g^{-1}, h^1, h^2, \ldots, h^{m-1}\}$ is a generator defied as follows:

- $f(a_1^{i_1}a_2^{i_2}.....a_n^{i_n}) = a_2^{i_2}.....a_n^{i_n}a_1^{i_1+1(\bmod 2m)}$;
- $f^{-1}(a_1^{i_1}a_2^{i_2}.....a_n^{i_n}) = a_n^{i_n-1(\bmod 2m)}a_2^{i_2}.....a_{n-1}^{i_{n-1}}$;
- $g(a_1^{i_1}a_2^{i_2}.....a_n^{i_n}) = a_2^{i_2}.....a_n^{i_n}a_1^{i_1}$;
- $g^{-1}(a_1^{i_1}a_2^{i_2}.....a_n^{i_n}) = a_n^{i_n}a_1^{i_1}.....a_{n-1}^{i_{n-1}}$;
- $h_j(a_1^{i_1}a_2^{i_2}.....a_n^{i_n}) = a_1^{i_1}.....a_{n-1}^{i_{n-1}}a_n^{i_n+2j(\bmod 2m)}$, j=1,2,…,m-1.

Fig. 1 illustrate a 4-degree Cayley WG_2^4.

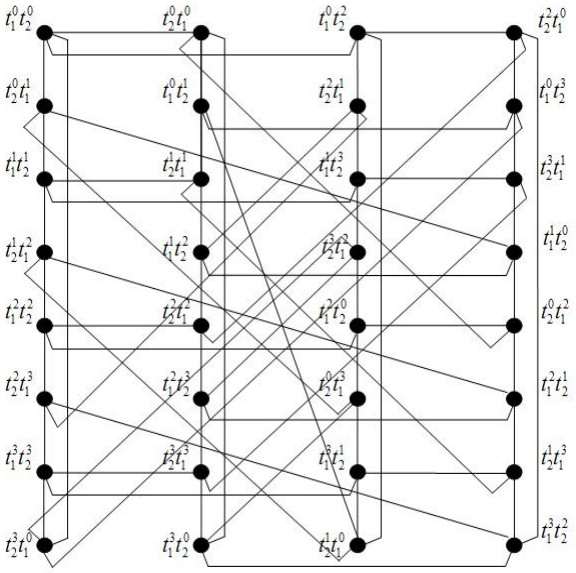


Figure 1. The Cayley graph WG_2^4

It has been verified that WG_n^2 is isomorphic to *wrapped butterlfly*[6], and its properties has been fully discussed in [19]. In this paper, we only consider graph WG_n^{2m} when $n\geq 2$ and $m\geq 2$.

The wreath product of the cyclic group Z_{2m} and symmetric group S_n, Z_{2m} *wr* S_n, is a group[20]. An element of the wreath product can be represented as $(g_1,g_2,\ldots,g_n;h)$, where $h\in S_n$ and $g_i\in Z_{2m}$, i=1,2,…,n; the multiplication in Z_{2m} *wr* S_n is:

$(g_1,g_2,\ldots,g_n;h)\cdot(g_1',g_2',\ldots,g_n';h')=(g_{h'(1)}g_1',\ldots,g_{h'(n)}g_n';hh')$

Let S={(0,0,…,0; 23…n1),

(0,0,…,0; n12….(n-1)),

(0,0,…,0,1;23…n1),

(2m-1,0,…,0;n12…(n-1)),

(0,0,…,0,2j;12…n)}, where $1\leq j\leq m$-1.

S is a generating set of Z_{2m} *wr* S_n, and $S=S^{-1}$. Then, we can define the Cayley graph $WG=Cay(Z_{2m}\ wr\ S_n, S)$. It is easy to verify that (0,0,…,0;123…n) is identity element.

Lemma 1. WG_n^{2m} ($n\geq 2$, $m\geq 2$) is a Cayley graph.

Proof. Let $V(WG_n^{2m})$ and $V(WG)$ denote the vertex set of WG_n^{2m} and WG, and a mapping φ from $V(WG_n^{2m})$ to $V(WG)$ is defined as follows:

φ: $a_1^{i_1}a_2^{i_2}.....a_n^{i_n} \rightarrow (g_1,g_2,\ldots,g_n;\ a_1'a_2'\ldots a_n')$, here $g_k=i_k$, and $a_i'=k_i$ *iff* $a_i=t_{k_i}$.

By the definition, φ is a bijection. Next, we show that φ preserves adjacency.

A node $u=a_1^{i_1}a_2^{i_2}.....a_n^{i_n}$ in WG_n^{2m} is adjacent to the node v. We shall show that their respective images in WG under φ are also adjacent vertices. For nodes u, $\varphi(u)=\varphi(a_1^{i_1}a_2^{i_2}.....a_n^{i_n})=(g_1,g_2,\ldots,g_n;\ a_1'a_2'\ldots a_n')$, because v is adjacent to u in WG_n^{2m}, we can give the derivation as follows:

Case 1. If $v=f(u)=a_2^{i_2}.....a_n^{i_n}a_1^{i_1+1 \bmod 2m}$, then

$\varphi(v)=\varphi(a_2^{i_2}.....a_n^{i_n}a_1^{i_1+1})$

$=(g_2,\ldots,g_n, g_1+1(mod\ 2m); a_2'\ldots a_n'\ a_1')$

$=(g_1,g_2,\ldots,g_n; a_1'a_2'\ldots a_n')$ (0,0,…,1; 23…n1)

$=\varphi(u)$ (0,0,…,1; 23…n1),

that is to say $(\varphi(u), \varphi(v))$ is an edge of WG.

Case 2. If $v=f^{-1}(u)=a_n^{i_n-1 \bmod 2m}a_2^{i_2}.....a_{n-1}^{i_{n-1}}$, then

$\varphi(v)=\varphi(a_n^{i_n-1}a_1^{i_1}a_2^{i_2}.....a_{n-1}^{i_{n-1}})$

$=(g_n\text{-}1, g_1,g_2,\ldots,g_{n-1}; a_n'a_1'a_2'\ldots a_{n-1}')$

$=(g_1,g_2,\ldots,g_n; a_1'a_2'\ldots a_n')$ (2m-1,0,…,0; n12…(n-1))

$=\varphi(u)$ (2m-1,0,…,1; n12…(n-1)),

that means $(\varphi(u), \varphi(v))$ is an edge of WG.

Case 3. If $v=g(u)=a_2^{i_2}.....a_{n-1}^{i_{n-1}}a_n^{i_n}a_1^{i_1}$, then

$\varphi(v)=\varphi(a_2^{i_2}.....a_{n-1}^{i_{n-1}}a_n^{i_n}a_1^{i_1})$

$=(g_2,\ldots, g_n, g_1; a_2'\ldots a_n'\ a_1')$

$=(g_1,g_2,\ldots,g_n; a_1'a_2'\ldots a_n')$ (0,0,…,0; 2…(n-1) n1)

$=\varphi(u)$ (0,0,…,0; 2…(n-1) n1),

that implies $(\varphi(u), \varphi(v))$ is an edge of WG.

Case 4. If $v=g^{-1}(u)=a_n^{i_n}a_1^{i_1}a_2^{i_2}.....a_{n-1}^{i_{n-1}}$, then

$\varphi(v)=\varphi(a_n^{i_n}a_1^{i_1}a_2^{i_2}.....a_{n-1}^{i_{n-1}})$

$=(g_n, g_1,g_2,\ldots,g_{n-1}; a_n'\ a_1'a_2'\ldots a_{n-1})$

$=(g_1,g_2,\ldots,g_n; a_1'a_2'\ldots a_n')$ (0,0,…,0; n12…(n-1))

$=\varphi(u)$ (0,0,…,0; n12…(n-1)),

Case 5. If $v=h_j(u)=a_1^{i_1}.....a_{n-1}^{i_{n-1}}a_n^{i_n+2j(\bmod 2m)}$, then

$\varphi(v)=\varphi(a_1^{i_1}.....a_{n-1}^{i_{n-1}}a_n^{i_n+2j(\bmod 2m)})$

$=(g_1,g_2,\ldots,g_n+2j(\ mod\ 2m); a_1'a_2'\ldots a_n')$

$=(g_1,g_2,\ldots,g_n; a_1'a_2'\ldots a_n')$ (0,0,…,2j; 12…(n-1) n)

$=\varphi(u)$ (0,0,…,2j; 12…(n-1) n),

that implies $(\varphi(u), \varphi(v))$ is an edge of WG.

According to **Case1-Case5**, we can show if u' and v' are adjacent in WG, their respective inverse images in WG_n^{2m} are also adjacent; thus we have WG and WG_n^{2m} are isomorphic. ■

According to the definition of WG_n^{2m}, we can easily get the following theorem.

Theorem 1. Cayley graph WG_n^{2m} ($n\geq 2$, $m\geq 2$):

(1) $d(WG_n^{2m})=m+3$ for $n>2$, and $d(WG_2^{2m})=m+2$;

(2) $|V(WG_n^{2m})|=n\cdot(2m)^n$;

(3) $|E(WG_n^{2m})|=n\cdot(2m)^n\cdot(m+3)/2$ for $n>2$, and

$|E(WG_2^{2m})|=n\cdot(2m)^n\cdot(m+2)/2$.

Theorem 2. Consider two graph models WG_n^{2m} and $G_{k,n}$, we have $\frac{d(WG_n^{2m})}{d(G_{k,n})}\approx\frac{1}{2}$ when $|V(WG_n^{2m})|=|V(G_{k,n})|$.

Proof. According to[9], the Cayley graph $G_{k,n}$ has $n(k\text{-}1)^n$ nodes. Then, we have $n\cdot(2m)^n=n(k\text{-}1)^n$, that is $m=(k\text{-}1)/2$. We can get $d(WG_n^{2m})=(k\text{-}1)/2+3$ for $n>2$, and $d(WG_2^{2m})=(k\text{-}1)/2+2$. Then,

$$\frac{d(WG_n^{2m})}{d(G_{k,n})}=\begin{cases}\dfrac{k-5}{2k} & \text{if } n>2;\\[2ex] \dfrac{k+3}{2k} & \text{if } n=2\end{cases}$$

Therefore $\frac{d(WG_n^{2m})}{d(G_{k,n})}\approx\frac{1}{2}$, when $k\to\infty$.■

III Routing Algorithm and Diameter

Since WG_n^{2m} is a Cayley graph, it is vertex symmetric, we can view the distance between any two arbitrary nodes as the distance between the source node and the identity permutation by suitably renaming the symbols representing the permutations. Thus in our subsequent discussion about a path from a source vertex to a destination vertex, the destination vertex is assumed to be the identity vertex $I=t_1^0t_2^0\ldots\ldots t_n^0$ without loss of generality. The following routing algorithm computes a path from an arbitrary source to the identity vertex I.

Definition 1. Consider an arbitrary node $v=a_1^{j_1}a_2^{j_2}\ldots\ldots a_n^{j_n}$ in WG_n^{2m}. Let $v[i]=j_i$, $1\le i\le n$, which denote the rank of the i^{th} symbol a_i in vertex v.

The Algorithm *RT*, which is shown in APPENDIX A, computes a path from an arbitrary source node $a_1^{j_1}a_2^{j_2}\ldots\ldots a_n^{j_n}$ in WG_n^{2m} to the identity node I.

Theorem 3. For an arbitrary node $a_1^{j_1}a_2^{j_2}\ldots\ldots a_n^{j_n}$ in WG_n^{2m}, the algorithm RT generates a path of length $\le\lfloor 5n/2\rfloor$.

Proof. If $k>\lceil n/2\rceil$ then $n\text{-}k+1$ $n-k+1\le\lfloor n/2\rfloor$ and hence after executing lines 1-20 of the algorithm, we reach the node $a_k^{j_k'}a_{k+1}^{j_{k+1}'}\ldots a_n^{j_n'}\ldots a_{k-1}^{j_{k-1}'}$ by a path of length less than $\lfloor n/2\rfloor$. Then, lines 21-36 are further executed, we reach the identity node by a path of length less than $2n$. Hence, we get the result. ■

IV Network Embeddings

In this section, we show the Cayley graph WG_n^{2m} can embed two kinds of structure of cycle and clique. A cycle structure is a sequence of distinct nodes $v_1,v_2,\ldots,v_p$ such that for each $i=1,2,\ldots,p\text{-}1$, (v_i,v_{i+1}) and (v_p,v_1) are in $E(G)$, and all edges (v_1,v_2), (v_2,v_3),…,(v_p,v_1) are distinct.

Definition 2. An edge introduced by the symmetric functions f or f^{-1} is called *f-edge*. A cycle in WG_n^{2m} consisting of only *f-edge* is called *f-cycle*. Similarly, an edge introduced by the symmetric functions g or g^{-1} is called *g-edge*. Any cycle in WG_n^{2m} consisting of only *g-edge* is called *g-cycle*. Consider an arbitrary node $v=a_1^{i_1}a_2^{i_2}\ldots\ldots a_n^{i_n}$ in WG_n^{2m}, then node v and nodes set $h_j(v)$, $1\le j\le m\text{-}1$, constitute a clique in WG_n^{2m} which is called a *h-clique*.

Definition 3. For each *f-cycle*, the unique node $v=t_1^0t_2^{i_2}\ldots\ldots t_n^{i_n}$ is called *f-leader*. Similarly, we can define the *g-leader* to be the unique node $v=t_1^{i_1}t_2^{i_2}\ldots\ldots t_n^{i_n}$ for each *g-cycle*. For each *h-clique*, the special node $a_1^{i_1}a_2^{i_2}\ldots\ldots a_{n-1}^{i_n-1}a_n^{*}$, where *=0 or 1, is called *h-leader*.

In the rest of this paper, we can use *f-leader* (respectively, *g-leader* and *h-leader*) to denote *f-cycle* (respectively, *g-cycle* and *h-clique*).Two *cycles* (*cliques*) are node-disjoint if they have no common node. For a nonnegative integer i, we recursively define

$$f^i(v)=\begin{cases}v & \text{if } i=0;\\ f(f^{i-1}(v)) & \text{if } i>0.\end{cases}$$

The notation $f^{-i}(v)$, $g^i(v)$, and $g^{-i}(v)$ can be defined similarly.

Theorem 4. WG_n^{2m} can embed $(2m)^{n-1}$ node-disjoint *f-cylces* of length $2m\cdot n$.

Proof. Given any node v in WG_n^{2m}, it is easy to verify that $f^{2mn}(v)=v$ and $f^i(v)\ne f^j(v)$, where $1\le i,j\le 2m\cdot n$ and $i\ne j$. Thus, from an arbitrary node v, a *f-cycle* of length $2m\cdot n$ can be generated by the functional iteration $f^{2mn}(v)$. Because WG_n^{2m} has $n(2m)^n$ nodes, there are $(2m)^{n-1}$ *f-cycles* each of length $2m\cdot n$ can be generated. These *f-cycles* are node-disjoint by the fact that $f(u)$-$f(v)$ if and only if $u=v$. ■

Definition 3.

(1) Two *f-cycles* f_1 and f_2 are adjacent if there exists a node $u\in f_1$ and a node $v\in f_2$ such that $u=\delta(v)$, where $\delta\in\{g,g^{-1},h_j\}$, and $1\le j\le m\text{-}1$.

(2) Two *g-cycles* g_1 and g_2 are adjacent if there exists a node $u\in g_1$ and a node $v\in g_2$ such that $u=\delta(v)$, where $\delta\in\{f,f^{-1},h_j\}$, and $1\le j\le m\text{-}1$.

(3) Two *h-cliques* h_1 and h_2 are adjacent if there exists a node $u\in h_1$ and a node $v\in h_2$ such that $u=\delta(v)$, where $\delta\in\{f,f^{-1},g,g^{-1}\}$.

Theorem 5. Each *f-cycle* of WG_n^{2m} is adjacent to $n(m+1)$ different *f-cycles* for $n>2$. In the case of $n=2$, any *f-cycle* of WG_n^{2m} is adjacent to $m+1$ different *f-cycles*.

Proof. Case 1. $n>2$.

For an arbitrary *f-cycle* F with *f-leader* $v=t_1^0t_2^{i_2}\ldots\ldots t_n^{i_n}$, we can get

(1) $(f^{-1})^l(h_j(f^t(v)))=t_1^0t_2^{i_2}..t_l^{i_l'}..t_n^{i_n}$, where $i_l'=(i_l+2j)(mod\ 2m)$ for all $l\ne 1$, and $1\le j\le m\text{-}1$.

Thus, $(n\text{-}1)(m\text{-}1)$ *f-leaders* of $(n\text{-}1)(m\text{-}1)$ distinct *f-cycles* can be obtained form v.

(2) $(f^{-1})^{1+2nj}(h_j(f(v)))=t_1^0 t_2^{i_2'}.....t_n^{i_n'}$, where $i_k'=(i_l-2j)(mod\ 2m)$ for all $1\le j\le m-1$.

Thus, *f-cycle F* is adjacent to $(m-1)$ distinct *f-cycles*.

(3) $f^{n-1}(g(v))=t_1^0 t_2^{i_2+1}.....t_n^{i_n+1}$.

(4) $(f^{-1})^{l-1}(g^{-1}(f^l(v)))=t_1^0 t_2^{i_2}..t_l^{i_l'}..t_n^{i_n}$, where $i_l'=(i_l+1)(mod\ 2m)$ for all $l\ge 2$.

The *f-cycle F* is adjacent to $(n-1)$ distinct *f-cycles*.

(5) $(f^{-1})^n(g^{-1}(f(v)))=t_1^0 t_2^{i_2-1}.....t_n^{i_n-1}$.

(6) $(f^{-1})^{l+1}(g(f^l(v)))=t_1^0 t_2^{i_2}..t_l^{i_l'}..t_n^{i_n}$, where $i_l'=(i_l-1)(mod\ 2m)$ for all $l\ge 1$.

The *f-cycle F* is adjacent to $(n-1)$ distinct *f-cycles*.

According to (1)-(6), *f-cycle F* is adjacent to $n(m+1)$ distinct *f-cycles*.

Case 2. $n=2$, in this case $g=g^{-1}$.

For an arbitrary *f-cycle F* with *f-leader* $v=t_1^0 t_2^{i_2}$, we can get

(1) $(f^{-1})^l(h_j(f^l(v)))=t_1^0 t_2^{i_2'}$, where $i_2'=(i_2+2j)(mod\ 2m)$ for all $l\ne 1$, and $1\le j\le m-1$..

Thus, $(m-1)$ *f-leaders* of $(m-1)$ distinct *f-cycles* can be obtained form *v*.

(2) $(f^{-1})^{l-1}(g(f^l(v)))=t_1^0 t_2^{i_2'}$, where $i_2'=(i_2+1)(mod\ 2)$ for all $l\ge 0$, and $l(mod\ 2)=0$.

(3) $(f^{-1})^{l+1}(g(f^l(v)))=t_1^0 t_2^{i_2'}$, where $i_2'=(i_2-1)(mod\ 2)$ for all $l\ge 1$, and $l(mod\ 2)\ne 0$.

According to (1)-(3), *f-cycle F* is adjacent to $(m+1)$ distinct *f-cycles*.■

Theorem 6. WG_n^{2m} can embed $(2m)^n$ node-disjoint *g-cylces* of length *n*.

Proof. Similar to Theorem 4. ■

Theorem 7. Each *g-cycle* of WG_n^{2m} is adjacent to $n(m+1)$ different *g-cycles*.

Proof. For an arbitrary *g-cycle C* with the *g-leader* $v=t_1^{i_1}t_2^{i_2}.....t_n^{i_n}$, we can get

(1) $(g^{-1})^{l+1}(h_j(g^l(v)))=t_1^{i_1}t_2^{i_2}..t_l^{i_l'}..t_n^{i_n}$, where $i_l'=(i_l+2j)(mod\ 2m)$ for $1\le j\le m-1$.

Thus, $n(m-1)$ *g-leaders* of $n(m-1)$ distinct *g-cycles* can be obtained form *v*.

(2) $(g^{-1})^{l+1}(f(g^l(v)))=t_1^{i_1}t_2^{i_2}..t_l^{i_l'}...t_n^{i_n}$, where $i_l'=(i_l+1)(mod\ 2m)$ for all $0\le l\le n-1$.

Thus, *g-cycle C* is adjacent to *n* distinct *g-cycles*.

(3) $(g^{-1})^{l-1}(f^{-1}(g^l(v)))=t_1^{i_1}t_2^{i_2}..t_l^{i_l'}..t_n^{i_n}$, where $i_l'=(i_l-1)(mod\ 2m)$ for all $1\le l\le n$.

Thus, *g-cycle C* is adjacent to *n* distinct *g-cycles*.

According to (1)-(3), *g-cycle C* is adjacent to $n(m+1)$ distinct *g-cycles*. ■

Theorem 8. WG_n^{2m} can embed $n(2m)^n/m$ node-disjoint *h-clique* of size *m*.

Proof. For an arbitrary node $v=t_1^{i_1}t_2^{i_2}.....t_n^{i_n}$, the node set $H=\{v\}\cup\{h_j(v)|\ 1\le j\le m-1\}$ constitute a *h-clique*, and $|H|=m$. It is obviously that there no common node between any two *h-cliques*. So, WG_n^{2m} can embed $n(2m)^n/m$ node-disjoint *h-clique* of size *m*. ■

Theorem 9. Each *h-clique* of WG_n^{2m} is adjacent to $4m$ different *h-cliques* for $n>2$. In the case of $n=2$, any *h-clique* of WG_2^{2m} is adjacent to $3m$ different *h-cliques*.

Proof. Case 1. $n>2$.

Given an arbitrary *h-clique* with *h-leader* $v=a_1^{i_1}a_2^{i_2}.....a_{n-1}^{i_n-1}a_n^*$, where *=0 or 1, we have

(1) $v'=f(v)=a_2^{i_2}.....a_{n-1}^{i_n-1}a_n^*a_1^{i_1+1}$, and v' is a node of *h-clique* H_1 with *h-leade* $a_2^{i_2}.....a_{n-1}^{i_n-1}a_n^*a_1^0$ when $i_1(mod\ 2)=1$ or $a_2^{i_2}.....a_{n-1}^{i_n-1}a_n^*a_1^1$ when $i_1(mod\ 2)=0$.

(2) $v'=g(v)=a_2^{i_2}.....a_{n-1}^{i_n-1}a_n^*a_1^{i_1}$, and v' is a node of *h-clique* H_2 with *h-leader* $a_2^{i_2}.....a_{n-1}^{i_n-1}a_n^*a_1^0$ when $i_1(mod\ 2)=0$ or $a_2^{i_2}.....a_{n-1}^{i_n-1}a_n^*a_1^1$ when $i_1(mod\ 2)=1$.

3) $v'=f^{-1}(v)=a_n^{*-1}a_1^{i_1}a_2^{i_2}.....a_{n-1}^{i_n-1}$, and v' is a node of an *h-clique* H_3 with *h-leader* $a_n^{*-1}a_1^{i_1}a_2^{i_2}.....a_{n-1}^0$ when $i_{n-1}(mod\ 2)=0$ or $a_n^{*-1}a_1^{i_1}a_2^{i_2}.....a_{n-1}^1$ when $i_{n-1}(mod\ 2)=1$.

(4) $v'=g^{-1}(v)=a_n^*a_1^{i_1}a_2^{i_2}.....a_{n-1}^{i_n-1}$, and v' is a node of an *h-clique* H_4 with *h-leader* $a_n^*a_1^{i_1}a_2^{i_2}.....a_{n-1}^0$ when $i_{n-1}(mod\ 2)=0$ or $a_n^*a_1^{i_1}a_2^{i_2}.....a_{n-1}^1$ when $i_{n-1}(mod\ 2)=1$.

Each *h-clique* of WG_n^{2m} is adjacent to $4m$ different *h-cliques*, because there are *m* nodes in it.

Case 2. $n=2$.

Similar to Case 1, note that $g^{-1}(v)=g(v)$.■

Table 1. Comparison of WG_n^{2m} and $G_{k,n}$

Network	No.nodes	Degree	Diameter
WG_5^6	$5\cdot 6^5$	6	12
WG_5^8	$5\cdot 8^5$	7	12
WG_5^{10}	$5\cdot 10^5$	8	12
WG_5^{12}	$5\cdot 12^5$	9	12
WG_5^{14}	$5\cdot 14^5$	10	12
WG_n^{2m}	$n(2m)^n$	$m+3$ or $m+2$	$\le\lfloor 5n/2\rfloor$
$G_{7,5}$	$5\cdot 6^5$	7	10
$G_{9,5}$	$5\cdot 8^5$	9	10
$G_{11,5}$	$5\cdot 10^5$	11	10
$G_{13,5}$	$5\cdot 12^5$	13	10
$G_{15,5}$	$5\cdot 14^5$	15	10
$G_{k,n}$	$n(k-1)^n$	k	$\le\lfloor 5n/2\rfloor$-2 or $2n$

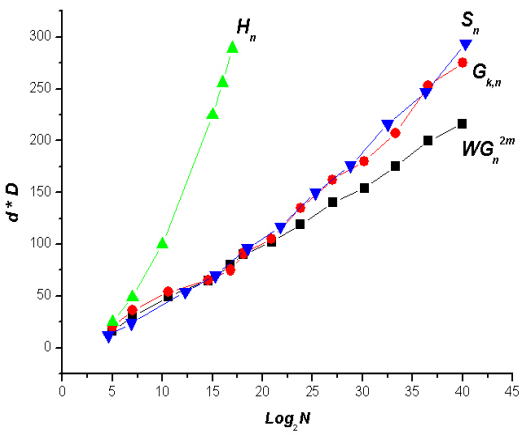


Figure.2. Comparision of the cost measure

V CONCLUSIONS

In this section, we consider some important parameters, such as node degree(d), diameter(D) and their product($d\times D$), namely cost measure. From Table 1, we can see that when WG_n^{2m} and $G_{k,n}$ have the same number of nodes and approximate diameter, the degree of WG_n^{2m} is much smaller.

APPENDIX A ALGORITHM *RT*

ALGORITHM RT(v)

compute k, $1\le k\le n$, such that $a_k=t_1$.
if $k>\lceil n/2\rceil$ then
  for i=1 to n-k+1 do
    if $v[n](mod\ 2)=0$ or $v[n]=-1$ then
      $v=g^{-1}(v)$
    endif
    if $v[n](mod\ 2)=1$ and $v[n]\neq -1$ then
      $v=f^{-1}(v)$
    end if
  end for
else
  for i=1 to k-1 do
    if $v[1](mod\ 2)=0$ or $v[1]=-1$ then
      $v=g(v)$
    endif
    if $v[1](mod\ 2)=1$ and $v[1]\neq -1$ then
      $v=f(v)$
    end if
  end for
end if
for i=1 to n do
  if $v[1]=-1$ then
    $v=f(v)$
  end if
  if $v[1]=0$ then
    $v=g(v)$
  end if
  if $v[1](mod\ 2)=0$ and $v[1]\neq 0$ then
    $v=g(v)$
    $v=h_{m-v[n]/2}(v)$
  end if
  if $v[1](mod\ 2)=1$ and $v[1]\neq -1$ then
    $v=f(v)$
    $v=h_{m-v[n]/2}(v)$
  end if
end for

ACKNOWLEDGMENT

This research partially supported by the Natural Science Foundation of China (60773083).

REFERENCES

[1] S.B. Akers, B.Krishnamurthy, A group-theoretic model for symmetric interconnection networks, IEEE Trans. Comput. 38 (4) (1989) 556-566.
[2] F.Annexstein, M. Baumslag, A.L. Rosenberg, Group action graphs and parallel architectures, SIAM J. Comput. 19(1990) 544-569.
[3] B.W. Arden, K.W. Tang, Representations and routing of Cayley graphs, IEEE Trans. Commun. 39 (11) (1991) 1533-1537.
[4] L. Bhuyan, D.P. Agrawal, Generalized hypercube and hyperbus structure for a computer network, IEEE Trans. 33 (1984) 323-333.
[5] N.L. Biggs, Algebraic Graph Theory, Cam bridge University Press, Camvradge, 1993.
[6] G.H.Chen, F.C.M. Lau, Comments on "A new family of Cayley graph interconnectionnetworks of constant degree four", IEEE Trans Parallel Distrib Syst 8(1997) 1299-1300.
[7] R.K.Das, B.P.Sinha, A new topology with odd degree for multiprocessor systems, J. Parallel Distrib. Comput. 39 (1996) 87-94.
[8] W.C. Fu, S.C. Chau, Cyclic-cubes: A new family of interconnection networks of even fixed-degrees, IEEE Trans. Parallel Distrib Syst 6 (1998), 1253-1268.
[9] S.Y. Hsieh, T.T. Hsiao, The k-Degree Cayley Graph and its Topological Properties, Networks 47 (2006), 26–36.
[10] S.Latifi, P.K.Srimani, A fixed degree regular network for massively paralled system, J.Supercompt. 12 (1998)
[11] S.Okawa, Correction to the diameter of trivalent Cayley graphs, IEICE Trans Fundam E84-A(2001), 1269-1272.
[12] F.Preparata, J.Vuillemin, The Cube-Connected cycles versatile network for parallel computation, Comm. ACM 24(5) (1981) 30-39.
[13] K. Qiu, S.G.Akl, H.Meijer, On some properties and algorithms for the star and pancake interconnection networks, J.Parall and Distributed Computing, 1993.
[14] C.H. Tsai, C.N.Hung, L.H.Hsu,C.H.Chang, The correct diameter of trivalent Cayley graphs, Inform Process Lett 72(1999), 109-111.
[15] P.Vadapalli, P.K.Srimani, Trivalent Cayley graphs for interconnection networks, Inform Process Lett 54 (1995), 329-335.
[16] P.Vadapalli, P.K.Srimani, A new family of Cayley graph interconnection networks of constant degree four, IEEE Trans Parallel Distrib Syst 7 (1996), 177-181.
[17] P.Vadapalli, P.K.Srimani, Shorstest routing in trivalent Cayley graph network, Infor Process Lett 57 (1996), 183-188.
[18] M.D.Wagh, J.mo, Hamilton cycles in trivalent Cayley graphs, Inform Process Lett 60 (1996), 177-181.
[19] D.S.L. Wei, F.P. Muga II, K.Naik, Isomorphism of degree four cayley graph and wapped butterfly and their optimal permutation routing algorithm, IEEE Trans Parallel Distrib Syst 10 (1999) 1290-1298.
[20] J. Xu, Topological structure and analysis of interconnection networks, Kluwer Academic Publishers, Dordrecht,2001.
[21] Chi-Hsiang Yeh, Emmanouel A. Varvarigo, Parallel algorithms on the Rotation-Exchange network-A trivalent variant of the star graph, Proceedings of the Symposium on Frontiers of Massively Paralle Computation vol. 1 (1999) 302-309.
[22] S.M. Zhou, N. Du, B.X. Chen, A new family of interconnection networks of odd fixed degrees.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 218-221

Optimizing Polynomial Window Functions by Enhanced Differential Evolution

Dongli Jia[1,2], Guoxin Zheng[2], Yazhou Zhu[2], and Li Zhang[2]
[1] Hebei University of Engineering / School of Information and Electronic Engineering , HanDan, China
Email: jwdsli@gmail.com
[2] Shanghai University / Key Laboratory of Special Fiber Optics and Optical Access Networks, ShangHai, China

***Abstract*—This paper proposed a class of polynomial window functions whose coefficients are optimized with the Enhanced Differential Evolution (EDE). Through optimization, the proposed window not only has an optimal solution, but can support different kinds of prescribed applications. The comparisons and analysis also show that the proposed method has a better performance than the traditional ones such as Hamming window, Hanning window, and the Blackman window.**

***Index Terms*—Differential Evolution, window functions, optimization**

I. INTRODUCTION

Window functions play an important role in traditional signal processing applications such as finite impulse response (FIR) filters design, spectrum analysis, speech recognition, and so on. Many window functions have been proposed in the literature [1-5]. All of these functions can be generally divided into two categories: fixed shape window and variable parameter window. The most commonly used fixed shape window functions include Rectangular window, Hanning window, Hamming window, Triangular window, and Blackman window etc. These windows are characterized by fixed spectral main-lobe width and only support specific applications. The other kind of window function, variable parameters window, includes Kaiser window, Gaussian window, and Chebyshev window. These window functions have at least one parameter which can be adjusted to form windows with different shapes and spectral features. For example, the Kaiser window derives from modified zeroth-order Bessel function with an adjustable parameter β . By controlling it different spectral features can be obtained such as the main-lobe width and the maximum side-lobe level to adapt to various applications.

However, all above window functions are suboptimal solutions and their uses heavily depend on the applications. Many researchers are still working on the area expecting to find the best windows. On the other hand, choosing a right window for a specific application is not an easy task. For example, in the FIR filters design, if the signal contains strong interfering frequency far from the frequency of interest, a window with a high side-lobe roll off rate should be chosen. However if the interfering signal is close to the frequency of interest, a window with a rather low maximum side-lobe level is a good choice. Further, if the interfering signal adjacent to the signal of interest, a window with a narrow main-lobe is more suitable.

To provide a better solution to the use of window functions, we proposed a class of optimized window functions with $w(t)$ defined by a sixth order polynomial of t:

$$w(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_5 t^5 + a_6 t^6 \quad (1)$$

Generally, the polynomial equation can form any window shapes to some extent if its order high enough. But in practice and for simplicity, we found the sixth order is adequate.

Differential evolution is a powerful heuristic search technique and wildly used to obtain the optimal values of objective functions. By means of the differential evolution, the coefficients $\{a_0, a_1, \cdots, a_6\}$ of the propsed window function can be optimized and its solutions are most close to the optimal values according to the prescribed spectral characteristics.

The proposed method has two main advantages: First, under the same condition, the DE method can get the best optimal solution compared with the windows already in the former literature. Second, the DE method can satisfy some special requirements to alleviate the choosing task.

The rest of the paper will focus on the implementation of Differential Evolution on window functions' optimizing. In section 2, the Enhanced Differential Evolution is briefly introduced. In section 3, the proposed method is depicted in detail. In section 4, the comparison with three traditional windows is given and its results are analyzed. Section 5 gives the conclusions.

II. ENHANCED DIFFERENTIAL EVOLUTION ALGORITHM

Differential Evolution (DE), which was first proposed over 1994-1996 by Storn and Price at Berkeley, is a very simple and very powerful evolutionary algorithm [6-7]. It has been proved that DE is an accurate, reasonably fast, and robust optimizer for function optimization in various fields such as filter design, PID control, image segmentation, and other scientific and engineering problems [8-9].

A. Standard DifferentialEevolution

Like other evolutionary algorithms, DE involves two phases: Initialization and evolution. In the first phase, the DE population is generated by random. In the second phase, individuals from the population go through

mutation, crossover, and selection process repeatedly until the termination criterion is met.

1. Initialization

DE is a parallel direct search technique using a population with N individuals for each generation g. The individual can be a vector:

$$X_i^g = [x_{i1}^g, x_{i2}^g, \cdots x_{ij}^g, \cdots, x_{in}^g] \quad (2)$$

where x_{ij}^g represents the *j*th potential parameter from i individual in *g*th generation. In initialization phase, all individuals from the whole population are initialized randomly with uniform probability distribution in its search space.

2. Mutation

There are different mutation methods in DE. In the paper, the simplest ones are employed. For each individual vector X_i^g, DE generates a mutated vector V_i^g based on the difference between two randomly selected individuals.

$$V_i^g = X_3^g + F \cdot (X_2^g - X_1^g) \quad (3)$$

where F is called scaling factor. X_3^g, X_2^g and X_1^g are randomly selected individuals from the population.

3. Crossover

The aim of the crossover is to generate the candidate child individuals. Different crossover schemes include one-point crossover, multi-point crossover, binomial crossover, and exponential crossover. Here, binomial crossover is implemented.

In this scheme, the candidate child vector is generated by the following equation:

$$u_{ij}^g = \begin{cases} x_{ij}^g & rand(j) \le CR \\ v_{ij}^g & rand(j) > CR \end{cases} \quad (4)$$

where u_{ij}^g is a parameter of candidate child individual U_i^g, and the v_{ij}^g is a parameter of V_i^g. CR is called crossover factor and limit to [0, 1].

4. Selection

Selection takes the competition mechanism. The candidate child U_i^g and the old individual X_i^g competed according to the fitness. The winner will have the chance to survive the next generation.

$$X_i^{g+1} = \begin{cases} X_i^g & f(X_i^g) > f(U_i^g) \\ U_i^g & f(U_i^g) > f(X_i^g) \end{cases} \quad (5)$$

The DE repeat above B, C, and D process until the termination condition is met. The final output is a candidate solution to $f(x)$.

B. Enhanced Differential Evolution

To further improve the performance of the standard DE, the Chaos Local Search (CSL) is incorporated into [10]. With the CSL technique, the DE is more robust with the better search ability. The CSL defined as:

$$X_i^{'(g)} = (1-\lambda) X_i^{(g)} + \lambda \beta_j \quad (6)$$

where $X_i^{'(g)}$ is a new vector of individual X_i^g in *g*th generation produced by chaotic local search, β_j is generated by logistic chaos iteration:

$$\beta_j^{k+1} = \mu \beta_j^k (1-\beta_j^k), k = 1, 2, \cdots; \quad (7)$$
$$\beta_j \in (0,1), \beta_j \ne 0.25, 0.5, 0.75$$

where β_j is the *j*th chaotic variable in *k*th iteration.

The λ is the shrinking scale given by:

$$\lambda = 1 - \left| \frac{g-1}{g} \right|^m \quad (8)$$

where m controls the shrinking speed.

The $X_i^{'(g)}$ is revalued and if its fitness is better than X_i^g, then the X_i^g is replaced by $X_i^{'(g)}$.

With "shrinking" strategy, DE has better performance in terms of the ability to find the global optimum.

III. Design Window Functions with Enhanced Differential Evolution

A. DE coding

In the paper, only symmetrical windows are considered. The proposed window $w(nT)$ with the length of N in the time domain has the following form:

$$w(nT) = \begin{cases} a_0 + a_1 nT + a_2 (nT)^2 + a_3 (nT)^3 + a_4 (nT)^4 \\ +a_5 (nT)^5 + a_6 (nT)^6; 0 \le n \le \frac{N-1}{2} \\ a_0 + a_1 kT + a_2 (kT)^2 + a_3 (kT)^3 + a_4 (kT)^4 \\ +a_5 (kT)^5 + a_6 (kT)^6; k = N - n; \frac{N-1}{2} \le n \le N \end{cases} \quad (9)$$

where T is the samping period.

In the case of optimized window functions design, we define the polynomial coefficients $\{a_0, a_1, \cdots, a_6\}$ as a DE individual.

$$X = \{a_0, a_1, \cdots, a_7\} \quad (10)$$

Through the evolutionary computation and fitness evaluation, the best individual X can be obtained to form different shapes and support different kinds of applications.

B. Fitness evluation

There are three important parameters to distinguish the performance of the window functions in the spectral domain [3]: Main-lobe width (M_w), maxmum side-lobe level (S_l), and the side-lobe roll off rate (S_r). These parameters can be defined as:

M_w: the width between the first zero-crossing to the left and the first zero-crossing to the right in the spectrum.

S_l: the maximum side-lobe level in dB, with respect to the main lobe peak.

S_r: the asymptotic decay rate of the side-lobe peaks in dB/octave.

However, the three parameters are not indepentent of each other. For example, a narrow main-lobe usually means a high maximum side-lobe level, and a low maximum side-lobe level means a low side-lobe roll off rate. Thus, the window choosing is a trade-off among the three parameters and the optimition is a multi-objective one in fact.

The fitness function can be define as:

$$F = w_1 \cdot \min(M_w) + w_2 \cdot \min(S_l) + w_3 \cdot \max(S_r) \tag{11}$$

where, the w_1, w_2 ,and w_3 are the weights with respect to the three parameters.

Ⅳ. WINDOW SPECTRUM COMPARISONS AND ANALYSIS

To test the preformance of the proposed method, we compare it with three typical window functions: hamming window, hanning window, and the blackman window. The related optimized coefficients are given in Table 1. In the comparison, we set the discrete window length N=100.

A. Comparison with Hamming and hanning window

Hamming and hanning window are widely used window functions with the $M_w = 8\pi / N$. The shape comparison is shown in Fig 1 and its performance comparison in spectral domain in terms of the maximum side-lobe level S_l and side-lobe roll-off rate is shown in Fig 2. The numerical results of Fig 2 are summarized in Table 2.

The Fig 1 illustrate that the proposed window has a similar shape to the hamming window. But the Fig 2 and the Table 2 demonstrate that although the proposed window and the hamming window have the same side-lobe roll-off rate S_r =−6dB, the proposed window has a better perrformance in terms of the maximum side-lobe suppression. Its S_l is about 3 dB lower than the hamming window's. The hanning window has the best S_r with the worst S_l among the three window functions.

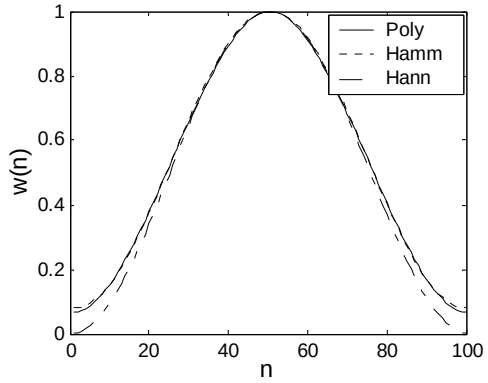


Figure 1. Window shapes comparison as $M_w = 8\pi / N$

B. Comparison with Blackman window

The spectrum of Blackman window has a wide main-lobe $12\pi / N$. It usually used in the occasion where side-lobe performance is more important. The comparison of the proposed window with the Blackman

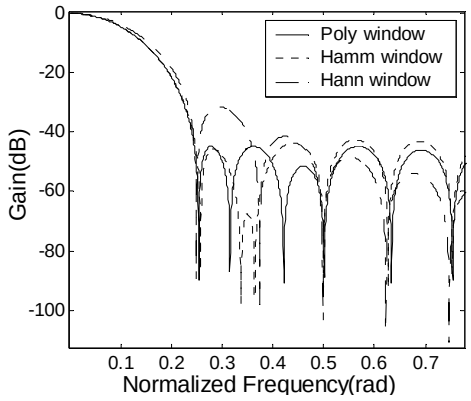


Figure 2. Spectrum comparison as $M_w = 8\pi / N$.

TABLE Ⅱ

NUMERICAL RESULTS COMPARISON AS $M_w = 8\pi / N$

Method	M_w (rad)	S_l (dB)	S_r (dB/oct)
Polynomial window	$8\pi / N$	-45.2	-6
Hanning window	$8\pi / N$	-31.6	-18
Hamming window	$8\pi / N$	-41.9	-6

TABLE Ⅰ

OPTIMIZED COEFFICIENTS OF PROPOSED WINDOW FUNCTIONS

M_w	a_0	a_1	a_2	a_3	a_4	a_5	a_6
$8\pi / N$	6.3138E-2	1.0021E-3	9.1112E-4	-5.6836E-6	-1.4142E-7	5.1242E-10	2.7636E-12
$12\pi / N$	5.7074E-3	-7.5151E-4	5.0199E-4	-1.9246E-5	1.7355E-6	-4.3014E-8	3.0575E-10

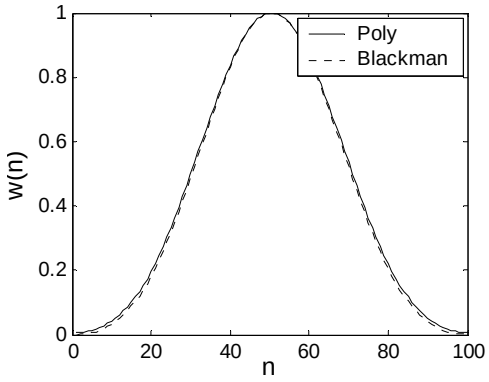


Figure 3. Window shapes comparison as $M_w = 12\pi / N$.

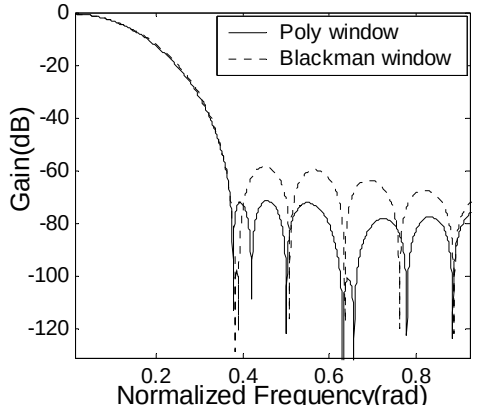


Figure 4. Spectrum comparison as $M_w = 12\pi / N$.

TABLE III

NUMERICAL RESULTS COMPARISON AS $M_w = 12\pi / N$

Method	M_w (rad)	S_l (dB)	S_r (dB/oct)
Polynomial window	$12\pi / N$	-71	-3
Blackman window	$12\pi / N$	-58	-18

window in terms of the parameters S_l and S_r is shown in Fig 3-4 and Table 3.

From the comparison results, it can be seen that the proposed window has a better performance in terms of the maximum side-lobe level. Its S_l is about 13 dB lower than the Blackman window's. But in terms of the parameter S_r , the Blackman window is better.

Ⅴ. CONCLUSIONS

This paper proposed a class of sixth order polynomial window functions. By means of the Enhanced Differential Evolution, the coefficients of polynomial window function can be optimized to obtain a better solution or support different applications. The comparisons and analysis show that the proposed method is better than the traditional window functions such as Hamming window and Blackman window.

ACKNOWLEDGMENT

This work was supported in part by Shanghai Leading Academic Discipline Project, STCSM （S30108 and 08DZ2231100）and NSFC （60872021）.

REFERENCES

[1] J.F. Kaiser, R.W. Schafer. "On the use of the IO-sinh window for spectrum analysis," *IEEE Trans on Acoustics Speech Signal Process.* vol. 28, pp:105–7, 1980.

[2] F.J. Harris. "On the use of window function for harmnic analysis with the discrete Fourier transform," *Proceedings of the IEEE*. Vol. 66, pp:51-83, Jan. 1978.

[3] K. Avcia, A. Nacaroglu. "Cosine Hyperbolic Window Family with its Application to FIR Filter Design," *3rd International Conference on Information and Communication Technologies: From Theory to Applications*. pp:1-6, April 2008.

[4] R.B. Blackman, J.W. Tukey. "The measurement of power spectra from the point of view of communication engineering," *Dover Publications.*

[5] SMA. Bergen, A. Antoniou. "Design of ultraspherical window functions with prescribed spectral characteristics," *EURASIP J Appl Signal Process.* vol. 13, pp:2053–65, 2004.

[6] R. Storn and K. Price. "Differential Evolution – A simple and efficient adaptive scheme for global optimization over continuous spaces," *Technical Report TR-95-012*, March 1995, ftp:ICSI.Berkeley.edu/pub/techreports/1995/tr-95-012.ps.

[7] R. Storn and K. Price, "Minimizing the real functions of the ICEC'96 contest by differential evolution," *Proc. of IEEE Int. Conf. on Evolutionary Computation*, Nagoya, Japan, 1996.

[8] R. Storn, Rainer. "Designing nonstandard filters with differential evolution", *IEEE Signal Processing Magazine.* vol.22, pp:103-106, January, 2005.

[9] Aslantas, Veysel, Tunckanat, Mehmet. "Differential evolution algorithm for segmentation of wound images," *IEEE International Symposium on Intelligent Signal Processing, WISP*, pp.1-5, 2007.

[10] Dongli Jia, Jiashu Zhang,"Niche particle swarm optimization combined with chaotic mutation," Control and Desion, Vol.22, No.1, Jan. 2007, pp.117-120.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 222-226

Gender Recognition with Face Images Based on PARCONE Model

Changqin Huang [1], Wei Pan [2], and Shu Lin [2]
[1] Department of Cognitive Science, Xiamen, Fujian, China
Email: cqinhuang@163.com
[2] Department of Cognitive Science, Xiamen, Fujian, China
Email: { wpan@xmu.edu.cn, linshu@ymail.com}

***Abstract*—In this paper, a new type of neural network model——PARCONE (Partially Connected Neural Evolutionary) was proposed, which can overcome the disadvantage that the previous neural networks can not accept more than thousands of inputs. With this new model, no feature extraction is needed before target identification and all of the pixels of a sample image can be used as the inputs of the neural network. After 300 ~ 600 generations' evolution, the new neural network can reach a good recognition rate. With this new model, a gender recognitionwe experiment was make on 490 face images (245 females and 245 males from Color FERET database), in which include not only frontal faces but also the faces rotated from -40°~40°in the direction of horizontal. The gender recognition rate, rejection rate and error rate of the positive examples respectively achieve 95.14%, 2.16% and 2.7%. The experimental results show that the new neural model has a strong pattern recognition ability and can be applied to many other pattern recognitions whick need a large amount of input information.**

***Index Terms*—neural network, PARCONE, face images, gender recognition rate**

I. INTRODUCTION

Gender classification is an important visual tasks for human beings, such as many social interactions critically depend on the correct gender perception. As visual surveillance and human-computer interaction technologies evolve, computer vision systems for gender classification will play an increasing important role in our lives.

As human faces provide important visual information for gender perception, a large number of researchers have investigated gender classification from human faces. In the early years, Moghaddam investigated nonlinear Support Vector Machines (SVM) for gender classification with low-resolution thumbnail face, and demonstrated the superior performance of SVM to other classifiers[1]. Walawalkar adopted SVMs for gender classification using audio and visual cues [2]. Shakhnarovich developed a real-time face detection and demographic analysis (female/male and asian/noasian) system using Adaboost, which delivers slightly better performance than the SVM on unaligned faces from real-world unconstrained video sequences [3].

Recently, various neural network techniques were employed for gender classification from human faces.[4-6]. Due to the size of neural network input vector with the increase of rapid growth, if we identify the image contain face with each pixel as the neural network input, it will make the neural network structure too complex to calculate the right output, and it will also cause badly real-time, or even non-convergence of network and other issues. Currently, researchers often use the methods of image feature extraction (such as border identification, principal components analysis (PCA), etc.). Then they train neural network with the features which are significantly reduced in the dimension of feature space. Despite this method could have been avoided excessive dimension, but to the specific issues, how to choose the characteristics? How many characteristics should be selected? There is not a unified approach. Using the eigenvector as the input of a neural network may be a simple way, but man-made feature selection would lose some of the key information of objectives, so the capacity of identification of neural network would be reduced.

It may be a contradiction that how to make the neural network to take full advantage of images containing important information, while not in a disaster of dimension. Our group proposed a new neural network model -- Model to solve this problem[8]. In the new model, firstly, every neuron can be connects with each other, but only a certain numbers connection values (for example 20) are not zero. Then we identify each image with every pixel value as a neural network input. During training, we change each neuron's previous 20 connections to other 20 connections by the rule of genetic algorithm (GA). For the sake of GA, we can gain the optimum (or sub optimum) 20 connections of every neuron. In this way, our neural network can automatically select the most important features of the objectives, achieve convergence and gain the capacity of target identification.

II. THE PARCONE MODEL

In earlier years, we evolved fully connected neural net modules (that all neurons can be connected with each other and are not limited by the connection number of a neuron.)[7], arguing that they were the "general case". By starting off with every possible connection (i.e. all N^2 of them if there are N neurons in the module) one could let the evolution decide if a particular connection should not exist, by driving down the value of its weighted connection to zero.

AP-PROC-CS-09CN005

This approach was fine, so long as the applications using the fully connected neural net modules did not require too many neurons *N*. If the sum of the pixel points of an image is more than thousands and all these pixels are inputted into the fully connected neural network, the evolvement time of the net will increase up to an unacceptable degree and the net may not achieve convergence. To deal with this problem, currently, there are two methods:

One of the methods is to reduce the size of the input images to an acceptable degree, such as hundreds of pixels, before inputting these images into the network. However, if the number of pixels falls below the 1000 mark, the quality of the image becomes rather poor and makes recognition between subtly different objects difficult. Another method is feature extraction (such as border identification, PCA, etc.), which has been discussed in the part of introduction.

Hence we decided to modify our old neural net model [8] which was fully connected, and make it partially connected.

The Partially Connected Neural Evolutionary (PARCONE) consists of three layers, as shown in Figure 1. They are:

1) Input layer, which has *I* neurons. Each of the neurons can input a pixel of a training image or a testing image. Each neuron of the input layer can be connected to *K* neurons of the Middle layer.

2) Middle layer, which contains *M* neurons. Each neuron of the Middle layer can be connected to *K* neurons of the Input layer, the Middle layer or the Output layer.

3) Output layer, which consists of *O* neurons, and each of the neurons can be connected to *K* Middle layer neurons.

I, *M*, *O* and *K* all can be change depended on the users, and $N=I+M+O$, where *N* is the sum of neurons of the whole net.

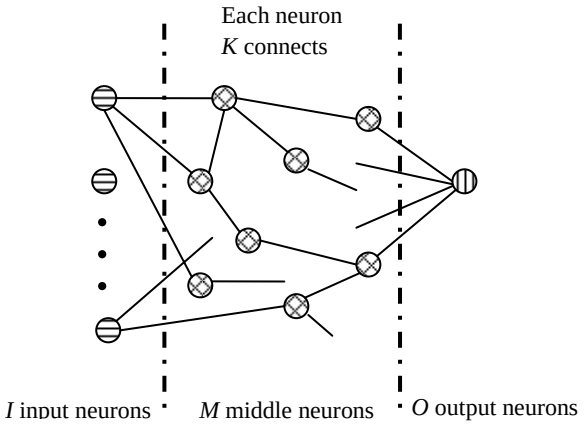


Figure 1. The structure of the new neuron network

Because there are so many neurons in our new network that the traditional standardization index-shaped function, such as:

$$S_i = (2.0/(1.0+e^{-A_i}))-1.0 \quad (1)$$

will be invalid. Our experiments show that function (2) can meet the needs of evolutional calculation of our new net [8].

$$S_i = A_i/(|A_i|+c) \quad (2)$$

where $A_i = \sum_{j=1}^{N} W_{ji} S_j$ is an usual active function of neural networks，*c* is a constant.

In PARCONE, for each neuron in the module, to list all the other neurons that that individual neuron connects to. Hence a partially connected neural net model will consist of a list of lists of inner-neural connections, one list per neuron. Each neuron in a module is given a unique non-negative integer ID.

That which specific neurons a neuron connects to and what its connection weight is are continuously adjusted according to the evolutionary algorithm. Using several entire original images (without any feature extraction) where include the objects to be identified as training samples, after 300~600 generations' evolutionary computations, each neuron will eventually be stably connected to *K* other neurons and *K* connecting weights occur, while the whole network acquires the ability to identify a specific object.

Figure 2 shows the data structures used in the coding of the Parcone model. A pointer points to a population of genetic algorithm chromosomes, i.e. pointers to a population of (partially connected) neural network modules. Each pointer in turn points to a further table of pointers, where each pointer points to a hash table for the neural net module in question. The hash table contains pointers to the structs with the ID, weight bits, and weight value. Thus the coding deals with pointers to a nested depth of 4, e.g.

To calculate the output signal of each neuron (at a given moment, called a "tick", where a "tick" is defined to be the time taken for all neurons in the module to calculate their neural output signals) in the module, its hash table is used. A scan across the length of the hash table is performed. From the previous "tick", a signal table of size N (the number of neurons in the module) is filled with the output signals of all the neurons. Assume the first non NULL pointer in the hash table is found at position (slot) "4". The non zero pointer there is used to find the "from" neuron ID. Let us say it is 45. The corresponding weight value might be 0.3452.

At initialization of the genetic algorithm, random connections between the all input neurons and the middle neurons are made, similarly between the middle neurons and other middle neurons or output neuron(s). Each input neuron is connected to a user specified number of middle neurons. Each output neuron is connected to a user specified number of middle neurons [8].

Evolution occurs by mutating the weight values, and/or by creating and deleting random connections.

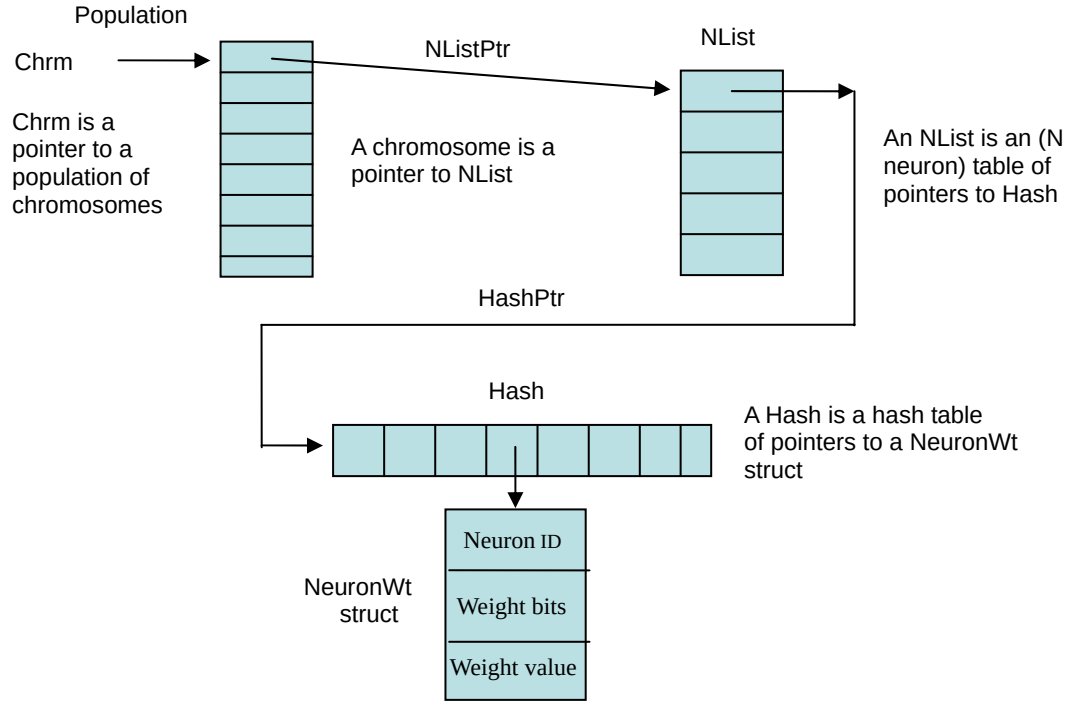
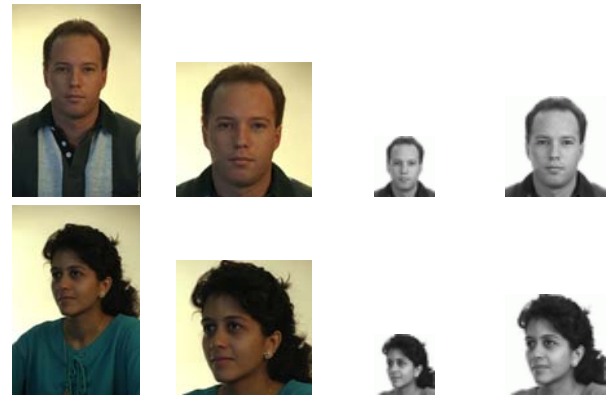


Figure 2. Data Structures of the Parcone Model



A: 512px×768px B: 432px×432px C: 30px×30px D: 50px×50px

Figure 3. The process of image pre-processing

III. EXPERIMENTAL RESULTS

Our experimental condition includes an ordinary PC machine (CPU: 2.40 GHz; Memory: 1 GB) and software Microsoft Visual C + + 6.0 and OpenCV 1.0.

A. Image Pre-processing

Our experiment uses total 490 face images (245 females and 245 males), randomly withdrawn from the Color FERET database. These images includes not only frontal faces but also the faces rotated from -40°~40°in the direction of horizontal(Figure 3A). Firstly, we cut out the raw images A (Figure 3A) and get images B (Figure 3B). Then, we convert B into gray images. After properly exposure adjusting and special size transformation of these gray images, we get images C (Figure 3C) and D (Figure 3D).

B. Experimental Results

In this paper, three groups of experiments are designed. The *Op* (output value of the neuron network) of each output neuron of our network is a decimal between -1 and 1.

To designate a male test sample as input of the network, when the output of the network $Op > 0.1$, we set that the identification of the network is correct; when $-0.1 < Op < 0.1$, we set that the network can not identify the input; when $Op < -0.1$, we set that the identification of the network is wrong.

To designate a female test sample as input of the network, when the output of the network $Op < -0.1$, we set that the identification of the network is correct; when $-0.1 < Op < 0.1$, we set that the network can not identify the input; when $Op > 0.1$, we set that the identification of the network is wrong.

Experiment 1: This experiment takes C-type images (30px×30px, Figure 3C) as training samples and testing samples. The training set includes 60 male images as positive examples and 60 female images as counter-examples, while the test set includes 185 male images and 185 female images.

The neural network (Figure 1) consists of three layers, the input layer, the middle layer and the output layer, in which there are I=900(neurons), M=600 (neurons) and O=1 (neuron)respectively. The number of connections of each neuron in the network is K(=20). The experimental results are as shown in table 1.

In table 1, P and N respectively denote positive example and counter-example.

TABLE 1.
THE EXPERIMENTAL RESULTS OF 30 × 30 SAMPLES (%)

Training generations		Recognition rate	Rejection rate	Error rate
300	P	81.62	4.86	13.52
	N	80.54	2.16	17.3
400	P	83.24	10.27	6.49
	N	77.3	3.24	19.46
500	P	69.19	7.03	23.78
	N	77.84	3.24	18.92
600	P	68.65	7.03	24.32
	N	79.46	2.16	18.38

Experiment 2: This experiment uses D-type images (50px ×50px, Figure 3D) as training samples and testing samples. The training set includes 60 male images as positive examples and 60 female images as counter-examples, while the test set includes 185 male images and 185 female images.

The neural network is as shown in Figure 1. The neuron number of the input layer, the middle layer and the output layer is 900, 600 and 1 respectively. The experimental results are as shown in table 2.

TABLE Ⅱ.
THE EXPERIMENTAL RESULTS OF 50 ×50 SAMPLES (%)

Training generations		Recognition rate	Rejection rate	Error rate
300	P	92.43	2.71	4.86
	N	72.97	3.78	23.25
400	P	95.14	2.16	2.7
	N	71.35	1.1	27.55
500	P	96.2	1.1	2.7
	N	71.35	2.16	26.49
600	P	94.06	2.16	3.78
	N	73.51	3.24	23.25

Experiment 3: In this experiment, the training set includes the training set of experiment 2 and the test samples which have not been correctly recognised in experiment 2, while the test set is the same as the test set of experiment 2.

TABLE Ⅲ.
THE EXPERIMENTAL RESULTS OF THE EXPERIMENT 3 (%)

Training generations		Recognition rate	Rejection rate	Error rate
300	P	88.11	3.24	8.65
	N	88.11	5.41	6.48
400	P	96.22	1.08	2.7
	N	92.43	1.62	5.95
500	P	92.43	1.62	5.95
	N	93.51	1.08	5.41
600	P	90.81	2.16	7.03
	N	93.51	2.16	3.23

The neural network is as shown in Figure 1. The neuron number of the input layer, the middle layer and the output layer is 900, 600 and 1 respectively. The experimental results are as shown in table 3.

C. Analysis of the experimental results

By the analysis of the experimental results, we get that:

1) Although we use all pixels of a sample image as the input (I = 2500) of the neural network, the novel Partially Connected Neural Evolutionary Model can still achieve convergence. It shows that this network can handle a large number of inputs.

2) The process of the image pre-processing done before the experiment is very simple. In the process, there is no feature extraction and almost all of the information of the original images has been retained. But in this way, not only the target information in the sample images, but also some other interference information, such as glasses, jewelry and different hairstyles, retains. That is why the recognition rates of the counter examples (female) are lower than the recognition rates of the positive examples (male) in experiment 1 and experiment 2.

3) That the face images we adopt for our experiments are not only frontal faces but also the faces rotated from -40°~ 40°in the direction of horizontal is different from the approaches shown in the previous literatures. On the one hand, it increases the difficulty of target identification. On the other hand, it makes our experiments more practical.

4) After 400 generations' evolution, the experimental results can preliminarily meet our expectation. For the sample images of 30px × 30px, the recognition rate of the positive examples is 83.24%, the rejection rate is 10.27% and the error rate is 6.49%, while for the sample images of 50px × 50px, the recognition rate, the rejection rate and the error rate of the positive examples are respectively 95.14%, 2.16% and 2.7%. The above experimental results show that the greater a sample image is, the more information it contains, and the more accurately our network identifies the targets. With the size increase of the sample images, our network can still achieve convergence, but it takes more time for evolution.

5) By adding the previous test samples which can not be identified or can not be identified correctly to previous training set, we get a bigger new training set. To evolve the new training set, the recognition rate can be improved, while the rejection rate and error rate are reduced. As table 3 shows, the recognition rate, the rejection rate and the error rate of the positive examples achieve 96.22%, 1.08% and 2.7% respectively.

6) The disadvantage of our experiment is that the training time is very long. It usually takes 10～30 hours for one experiment. To solve this problem, our group puts forward two possible solutions. One of the solutions is to use computer cluster. Another solution is to use a new introduced tool, Tesla S1070 (960 nuclear calculation module), which has been installed on our LAB. With the

new tool, the evolution speed of our neural network is expected to be increased about 200 times and an experiment can be completed in minutes.

IV CONCLUSIONS

Our objective is gender recognition with face images based on PARCONE model. The PARCONE model is a novel neural network and was first proposed and completed by our group. The results of experiments on human faces rotated from -40°~40°in the direction of horizontal show that our method is more feasible. But a lot of work should be done in the future. For example, how many neurons of each layer of the network is more appropriate, how to effectively improve the evolution speed and how many connections of each neuron in the network will be better.

ACKNOWLEDGMENT

The authors would like to thank Hugo de Garis for theoretic consultation and the Chinese National Natural Science Foundation(60975084) and the Science Foundation of Fujian Province of China(2009J01305) for financial support.

REFERENCES

[1] B.Moghaddam and M.Yang. Learning Gender with Support Faces[J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 2002,24(5):707–711.

[2] L.Walawalkar, M.Yeasin, A.Narasimhamurthy, R.Sharma. Support vector learning for gender classification using audio and visual cues[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2003, 17(3): 417–439.

[3] G.Shakhnarovich, P.A.Viola, B.Moghaddam. A Unified Learning Framework for Real Time Face Detection and Classification[C]// Proc. IEEE International Conference on Automatic Face and Gesture Recognition,2002:14–21.

[4] Tolba A S. Invariant gender identification[J]. IEEE Trans. Digital Signal Processing, 2001,11(3):222–240.

[5] Rowley H A, Baluja S, Kanade T. Neural Network-Based Face Detection[J]. IEEE Trans Pattern Analysis and Machine Intelligence,1998,201, 20(1) :25–38.

[6] Surendra Ranganath, Krishnamurthy Arun. Face recognition using transform features and neural networks [J]. PatternRecognition,1997,3010,30(10):1615– 1622.

[7] Hugo de GARIS. A "PARtially COnnected Neural Evolutionary" Model Serving as the Basis for Building China's First Artificial Brain[C]// Proceedings of 2008 3rd International Conference on Intelligent System and Knowledge Engineering. China: Xiamen, 2008:9–12.

[8] Hugo de Garis,Michael Korkin. The CAM-Brain Machine (CBM) An FPGA Based Hardware Tool which Evolves a 1000 Neuron Net Circuit Module in Seconds and Updates a 75 Million Neuron Artificial Brain f or Real Time Robot Control[J]. *Neurocomputing*, 2002, 42:35–68.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 227-230

Service Oriented Enterprise Application Integration and its Implementation Based on Open Source Software

Dongjin Yu[1, 2], and Guangming Wang[3]
[1] School of Business Administration, Zhejiang Gongshang University, Hangzhou, China
[2] Hangzhou Dianzi University, Hangzhou, China
Email: yudj@hdu.edu.cn
[3] Zhejiang Gongshang University, Hangzhou, China
Email: gmwang@mail.zjgsu.edu.cn

***Abstract*—The technology of Enterprise Application Integration could be applied for the interoperation among distributed heterogeneous systems. This paper presents the application integration framework for data exchange and business interaction based on Service Component Architecture, Message Oriented Middleware and Enterprise Service Bus. Furthermore, it introduces the implementation of the framework using some leading open source software such as Apache Tuscany, Apache ServiceMix and Apache ActiveMQ. The case conducted in the integration of 38 regional labor management information systems shows the framework is reliable and also has good performance with reduced cost.**

***Index Terms*—Enterprise Application Integration, Open Source, Service Oriented Architecture, Framework, Labor Management Information Systems**

I. INTRODUCTION

Enterprise Application Integration (EAI) is the process of linking different applications together within a single organization or across organization boundaries in order to simplify and automate business processes to the greatest extent possible, while at the same time avoiding having to make changes to the existing applications or data structures. In the words of the Gartner Group, EAI is the unrestricted sharing of data and business processes among any connected application or data sources in the enterprise [1]. EAI usually involves the data exchange which achieves a uniform data view of participating systems, and the business interaction which accomplishes mutual invocation across boundaries of autonomous systems in real time.

Service-oriented computing promotes the idea of assembling application components into a network of services that can be loosely coupled to create flexible, dynamic business processes and agile applications that span organizations and computing platforms. Here, services are referred as autonomous, platform-independent entities that can be described, published, discovered, and loosely coupled in novel ways [2]. Service-oriented solution is becoming a new approach to EAI whose framework could be constructed through universally accepted standards such as SCA, BEPL and WSDL.

As for the construction of EAI framework based on service oriented technology, the academic circle has already conducted a great deal of researching work. Related topics include the integration methodology or patterns, semantic service model, and so on. For instance, Zhang GS et al. present a formal systematic analysis, verification and validation methodology called SOARM. Based on Petri nets and temporal logic, SOARM could be well suited in EAI [3]. The idea of executable EAI patterns, which Scheibler et al. introduce in [4], is also worth attention. Using workflows customized with these configurable EAI patterns in a software-as-a-service (SaaS) business model, companies could focus on the integration without the need for the setup of complex integration infrastructures.

Most of current service-based EAI solutions adopt WSDL as the description of web service. Therefore the retrieval of right services relied on keywords will probably lead to unsuitable results with formally match but semantically not. To manage semantic differences, many propose the semantically described services for EAI. In [5], Liyi Zhang and Si Zhou give a framework of semantic SOA for EAI, and use Web Service Modeling Ontology (WSMO) as its semantic service model. Similarly in [6], Martinek et al. give an example about how to increase the effectiveness of integration by applying semantically described services.

Meanwhile, many real cases of service-based EAI have been developed. O. R. Bagheri et al. present the elastic EAI framework which has a service-based architecture and could be developed from the bottom up by means of existing technology [7]. Frequent mentioned service-based EAI cases usually occur in the fields such as Enterprise Information Portal [8], dynamic integration of Supply Chain [9], simply because these fields usually involve disparate heterogeneous systems, and more importantly, the integration in those fields could bring about huge profits.

Different with above mentioned ones, this paper gives a novel approach to service-oriented EAI framework based on open source software. With the underlying Service Component Architecture (SCA) implemented by Apache Tuscany, Message Oriented Middleware (MOM) implemented by ActiveMQ, and the Enterprise Service Bus (ESB) implemented by Apache ServiceMix, the framework fulfills both business interaction and data exchange with a variety of binding mechanisms such as Web Services, JMS and JCA, but needs less Total Cost of

Ownership (TCO) compared with using commercial products.

Service Component Architecture (SCA) provides a programming model for building applications and solutions based on a Service Oriented Architecture (SOA). It aims to encompass a wide range of technologies for service components and for the access methods which are used to connect them. In SCA, a component consists of a configured instance of an implementation, which offers services used by other components. In other words, implementations may depend on services provided by other components – these dependencies are called references [10].

Enterprise service bus (ESB) consists of a software architecture construct which provides fundamental services for complex architectures via an event-driven and standards-based messaging-engine (the bus) [11]. Unlike the more classical EAI approach of a monolithic stack in a hub and spoke architecture, the participants in the ESB-involved EAI framework do not need to interact with each other directly. Instead, the bus is responsible to deliver the request to the specific qualified service provider.

Message-oriented middleware (MOM) is the traditional infrastructure focused on sending and receiving messages that increases the interoperability of an application by allowing the application to be distributed over heterogeneous platforms [12]. MOM typically supports asynchronous calls between the client and server by the message queues which provide temporary storage when the destination program is busy or not connected.

The rest of the paper is organized in the following manner. Section 2 presents the architectural design for the service-oriented EAI framework. Section 3 illustrates its implementation based on open source software. The successfully implemented case is illustrated in Section 4. Finally, Section 5 provides concluding remarks and offers future research directions.

II. ARCHITECTURE OF SERVICE-BASED EAI FRAMEWORK

The core of service-based EAI framework is the Enterprise Service Bus configured with two kinds of adaptors to provide the interfaces of Web Services and messaging services respectively. The framework could also be extended to implement interfaces of JCA components or EJB components when required. In addition, the MOM and SCA components are bound to the bus via adaptors. The framework's architectural topology is illustrated in Fig. 1.

A. Design for business interaction

The framework is configured with the SCA-compliant process service engine above the Enterprise Service Bus to fulfill the business interaction across system boundaries. All existing shareable function units should be reconstructed as the standard service components. The interaction therefore could be realized between the coherent interfaces of services and references. Besides, the component implemented in BPEL is configured as the portal, which orchestrates the individual service components to achieve the business value. Finally, a dedicated data repository is indispensable, where the uniform master data and shareable business status data are kept.

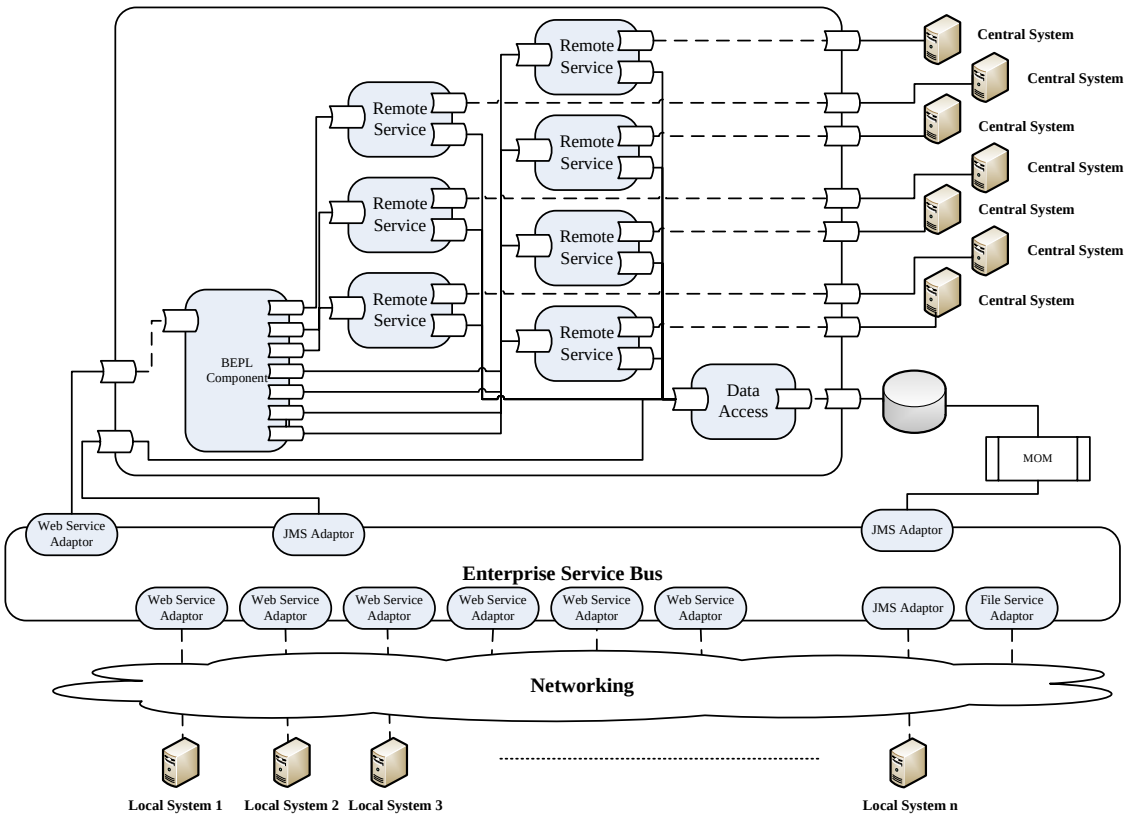


Figure 1. Service-based EAI framework

The business interaction in real scenarios would happen in the following manner. Each independent system accomplishes its own business process. When interaction across system boundaries required, a service request is issued and dispatched to the service bus. After the necessary transformation of communication protocols and data formats, the bus then routes the request to the portal component implemented in BPEL. The latter invokes corresponding service components in different systems and returns the response to the bus. The bus eventually delivers the result to the original requestor after reversal transformations of communication protocols and data formats.

B. Design for Data Exchange

The data exchange among different systems is inevitable during the process of Enterprise Application Integration. For example, considerable quantities of data would probably be migrated from one system to another after a fixed period of time. Under other circumstances, data representing business status in one system are required to keep synchronized with that on other systems simultaneously [13].

The framework realizes the data exchange based on messages. The configured Message Oriented Middleware is coupled with service bus through the specific adaptor. It maps the heterogeneous data formats, interfaces and protocols into the uniform ones, and provides reliable message filtering, pre-processing and transferring.

The framework provides two data synchronizing modes, named as Publish-Subscribe and Request-Response respectively.

1) Publish-Subscribe Mode

The data provider does not need to send the data directly to the receiver. Instead, the data are published under certain topics. The Message Oriented Middleware will propagate the data to the receiver subscribing it via the service bus. In this mode, the sending of data is initiated by the data provider.

2) Request-Response Mode

The node demanding data requests the Message Oriented Middleware for the data access service. The latter then interacts with the corresponding data provider via the service bus. Once the data obtained, the Message Oriented Middleware sends it back to the demanding node. In this mode, the sending of data is initiated by the demand side.

III. IMPLEMENTATION BASED ON OPEN SOURCE SOFTWARE

To construct the service-based EAI framework presented in this paper, different supporting platforms, or containers, are needed. The Service Component Architecture engine could adopt IBM Websphere Process Server or AquaLogic Data Services Platform. The Message Oriented Middleware could adopt Microsoft MSMQ or IBM Websphere MQSeries. The Enterprise Service Bus could adopt Iona Artix. However, the implementation with the above commercial products significantly increases the Total Cost of Ownership (TCO) due to the expensive license fee.

Therefore, the framework here integrates some leading open source software such as Apache Tuscany, Apache ServiceMix and Apache ActiveMQ, given in Table 1.

IV. CASE STUDIES

The above service-oriented EAI framework has been successfully implemented in the field of labor and social security administration. For this case, there are altogether 38 independent heterogeneous systems across 12 cities and counties, covering 7 lines of business including labor supervision, labor contract management, rural labor force management, employment management, unemployment management, vocational introduction and social security management. Some systems are maintained respectively by each city or county, while others provide services for the whole area.

With the emergence of frequent migration of labor forces, the demands for cross regional affairs such as remote job application, transferring of social insurance accounts become increasingly urgent. Accordingly, the service-oriented Enterprise Application Integration framework was introduced to accomplish these cross regional affairs and to monitor the demand and the required of labor forces in the whole area.

TABLE I. OPEN SOURCE SOFTWARE ADOPTED IN THE FRAMEWORK

	SCA Container	ESB Container	MOM
Software	Apache Tuscany	Apache ServiceMix	Apache ActiveMQ
Main Features	1) implements SCA Assembly Model V1.1 2) offers various component types including Java, C++, BPEL, Spring and scripting 3) provides a wide range of pluggable binding protocols such as RMI, Web Services, JSONRPC and EJB	1) built from the ground up on the JBI specification JSR 208 2) completely integrated into Apache Geronimo, JBoss and JOnAS	1) supports a variety of Cross Language Clients and Protocols from Java, C, C++, C#, Ruby, Perl, Python and PHP 2) supports JMS 1.1 and J2EE 1.4
Web Site	http://tuscany.apache.org	http://servicemix.apache.org	http://activemq.apache.org
Licensing	Apache License	Apache License	Apache License

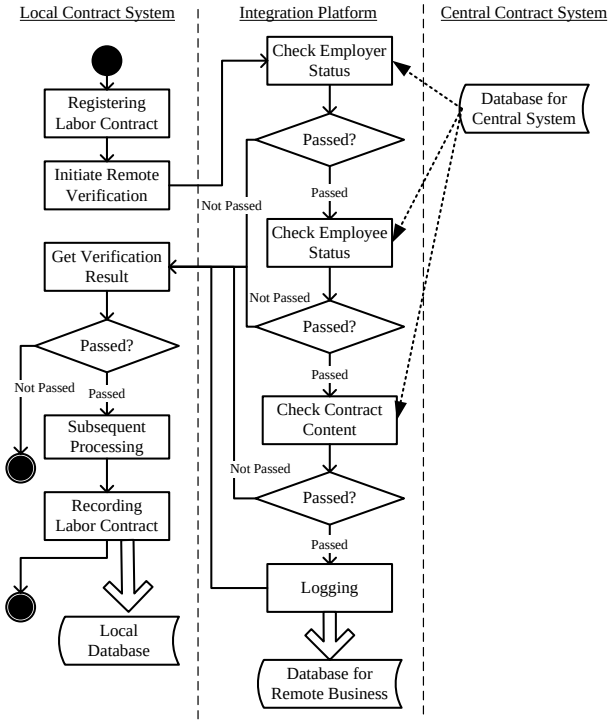


Figure 2. The process of labor contract verification

As the demo case, the labor contract verification process is illustrated in Fig. 2, where the Integration Platform is based on the above described service-oriented Enterprise Application Integration framework. In Fig. 2, the local labor contract management system first connects the corresponding Web service running in the Integration Platform, which then accesses the central database to check the status of both employers and employees. If the verification fails, failure reasons are returned and the process terminates. Otherwise, the Integration Platform checks the contents of contract. If successful, the Integration Platform backups the contract information in the local data repository and finally sends the messages back to the local system.

The Integrated Platform has already coupled 4 separate local labor contract systems in four cities after it started up in 2008. On average, 4000 till 5000 requests of labor contract verification are handled through the Integrated Platform each day. The response time generally does not exceed one second during the peak period.

V. CONCLUSIONS

Modeling the processes of data exchange and business interaction among loosely coupled heterogeneous systems based on services meets the demands of flexibility and reusability during the process of Enterprise Application Integration. Service Component Architecture, Message Oriented Middleware and Enterprise Service Bus are three cornerstones of service-based Enterprise Application Integration framework. The adoption of open source software such as Apache Tuscany, Apache AcitveMQ and Apache ServiceMix significantly reduces the total cost of ownership of service-based Enterprise Application Integration framework. Future work may include the development of its integrated management utilities which are expected to facilitate the end-users.

ACKNOWLEDGMENT

The work is supported by the Foundation of Zhejiang Provincial Key Science and Technology Projects (No. 2008C11099-1), and China Innovation Foundation for Technology-based Firms (No. 08C26213300677).

REFERENCES

[1] Gable, and Julie, “Enterprise application integration”, *Information Management Journal*, March/April, 2002.

[2] Papazoglou M. P., Traverso P., Dustdar S., and Leymann F., “Service-Oriented Computing: State of the Art and Research Challenges”, *IEEE Computer*, November 2007, pp. 64-71.

[3] Zhang GS, Jiang CJ, Tang XF, and Xu Y., “Specification and verification of service–oriented enterprise application integration system”, *Journal of Software*, 2007, 18(12), pp. 3015-3033.

[4] Scheibler T., Mietzner R., and Leymann F., “EAI as a Service - Combining the Power of Executable EAI Patterns and SaaS”, EDOC '08, Sept. 2008, pp. 107–116.

[5] Zhang Liyi, Zhou Si, and Zhu Mingzhu, “A Semantic Service Oriented Architecture for Enterprise Application Integration”, ISECS '09, 2009, pp. 102–106.

[6] Martinek P., Tothfalussy B., and Szikora B., “Semantically described services in the Enterprise Application Integration”, 30th International Spring Seminar on Electronics Technology, 2007, pp. 335–338.

[7] O. R. Bagheri, R. Nasiri, M. H. Peyravi, and P. Khosraviyan Dehkordi, “Toward an elastic service based framework for Enterprise Application Integration”, Fifth International Conference on Software Engineering Research, Management and Applications, 2007, pp. 711–719.

[8] Tang Xiao-xin, and Chen Guo-hua, “Construction of Enterprise Information Portal in Cigarette Enterprise”, International Conference of Management Science and Engineering, 2007, pp. 293–296.

[9] Chen Tingbin, Wang Lina, Zhang Yimin, and Sun Fuquan, “Research on Methods of Services-Oriented Integration for Supply Chain Collaboration”, WiCom 2007, pp. 4706-4709.

[10] Michael Beisiegel, et al., “SCA Service Component Architecture Assembly Model Specification: SCA Version 1.00”, 2007, http://www.osoa.org/download/attachments/35/SCA_AssemblyModel_V100.pdf?version=1.

[11] Schmidt Marc-Thomas, Hutchison Beth, Lambros Peter, et al., “The enterprise service bus: Making service-oriented architecture real”, *IBM Systems Journal*, 2005, 44(4), pp. 781-797.

[12] Sushant Goel, Hema Sharda, and David Taniar, “Message-Oriented-Middleware in a Distributed Environment”, IICS 2003, pp. 93-103.

[13] Dongjin Yu, Xindong You, and Quan Wang, “The Study and its Implementation of Resolving Data Synchronization Conflicts Based on Confidence Value”, Fourth International Conference on Computer Sciences and Convergence Information Technology, 2009, pp. 728-732.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 231-234

Adaptive Fuzzy Sliding Mode Control for Inverted Pendulum

Wu Wang
Xuchang University , School of Electrical and information Engineering, Xuchang, China
e-mail: jhwlz@tom.com

***Abstract*—A nonlinear sliding mode control method is presented for single inverted pendulum position tracking control. Sliding mode control (SMC) is a special nonlinear control method which have quick response, insensitive to parameters variation and disturbance, online identification for plants are not needed, its very suitable for nonlinear system control, but in reality usage, the chattering reduction and elimination is key problem in SMC. By using a function-augmented sliding hyperplane, it is guaranteed that the output tracking error converges to zero in finite time which can be set arbitrarily. Fuzzy logic systems are used to approximate the unknown system functions and switch item. Robust adaptive law is proposed to reduce approximation errors between true nonlinear functions and fuzzy systems. the definition of sliding mode control was presented and on the basis of the inverted pendulum system, the sliding mode controller was designed, Stability of the proposed control scheme is proved by Lyapnouv theorem and the control scheme is applied to an linear system and inverted pendulum system respectively, simulation studies shows the methods is effective and can applied into linear or nonlinear control system.**

***Index Terms*—adaptive fuzzy control; sliding mode control; inveted pendulumr; nonlinear system; simulation**

I. INTRODUCTION

An inverted pendulum system is a static unstable system, inverted pendulum has become a hot topic in control field for the similarity in control of helicopter launching of space shuttle and operation of satellite and robot walking with two legs [1]. Various control methodologies have been proposed for inverted pendulum systems and overhead trolley crane system they have things in common in the past. In the nonlinear system of inverted pendulum is linearized at zero and a popular pole placement approach is used to design a state feedback controller. Sliding mode control (SMC) is a special nonlinear control, since the publication of the pioneering paper on sliding mode control, significant results and many applications have been reported in the literature, SMC systems exhibit superb control performance which can be designed and have no relationship with controlled plant parameters and disturbance, also it have some advantages such as quick response, insensitive to parameters variation and disturbance, online identification for plants are not needed, but chattering reduction and elimination is key problem in SMC [2]. In basic SMC, big switching gain was needed to eliminate disturbance and uncertain factors, which was the main reason of chattering. Fuzzy control has many advantages such as control with mathematical models not needed, can use expert information and knowledge, and with strong robustness, on the other hand, in practical control equipment the controller parameters are hard to get and lack of systematic analytical method, Fuzzy sliding mode control (FSMC) combine fuzzy control with sliding mode control, FSMC make control destination from trace error to sliding mode function, if make sliding function to zero, the error can to zero gradually, for high order system, FMSC can hold two dimension input. FSMC is a soft control and can make chattering reduction and elimination. [3]. O. P. Ha applied equal control switching control and fuzzy control to realize fuzzy sliding mode controller [4], K. Y. Zhuang use fuzzy control estimate uncertain of system [5], S. H. Ryu design fuzzy rules based on chattering reduction [6], S. W. Kim divide sliding mode surface with fuzzy theory [7], B. Yoo approach unknown function with fuzzy system [8], C. Y. Liang design sliding mode surface with integral sliding mode function [9], J.Y. Chen use membership function adjust switching gain [10], P.G. Grossimon [11]and Y. P. Chen used SMC into inverted pendulum control [12].

II. ADAPTIVE FUZZY SLIDING MODE CONTROL

Definition of SMC

SMC is a control strategy of variable structure control, it's a special nonlinear control and mainly expressed as discontinuous [13].

Assume a control system can be described as:

$$\dot{x} = f(x,u,t) \qquad x \in \mathbf{R}^n, u \in \mathbf{R}^m, t \in \mathbf{R} \tag{1}$$

Switching function can be given:

$$s(x), \qquad s \in \mathbf{R}^m \tag{2}$$

Control function can be given:

$$u = \begin{cases} u^+(x) & s(x) > 0 \\ u^-(x) & s(x) < 0 \end{cases} \tag{3}$$

If the sliding mode exist and reachability condition satisfied, ie. Accepted for $s(x) = 0$, all movement spot can reach sliding mode surface and the sliding mode is stable, this is called sliding mode control [14].

Design of control system

Consider a SISO control system as follow [15]:

$$\ddot{\theta} = f(\theta,t) + g(\theta,t)u(t) + d(t) \tag{4}$$

Where $f(\theta,t)$ and $g(\theta,t)$ are unknown nonlinear function and $d(t)$ is external disturbance [16].

The trace error can be given:

AP-PROC-CS-09CN005

$$e(t) = \theta(t) - r(t) \tag{5}$$

The integral sliding mode surface can be defined as:

$$s(t) = \dot{\theta}(t) - \int_0^t [\ddot{r}(t) - k_1\dot{e}(t) - k_2 e(t)]dt \tag{6}$$

In (6), $k_1 > 0, k_2 > 0$.

If SMC in ideal state, $s(t) = \dot{s}(t) = 0$

So:

$$\ddot{e}(t) + k_1\dot{e}(t) + k_2 e(t) = 0 \tag{7}$$

Switching function $s(t)$ as input and the fuzzy rules can be given:

$$\text{Rule } i\text{: IF } s \text{ is } F_s^i \text{ THEN } u \text{ is } \alpha_i \tag{8}$$

$k_1 > 0, k_2 > 0$

With the center of gravity defuzzification method the output of controller can be given:

$$u_{fz}(s) = \frac{\sum_{i=1}^{m} \omega_i \alpha_i}{\sum_{i=1}^{m} \omega_i} \tag{9}$$

Where ω_i is the weight of according Rule i.

Assume $f(\theta,t)$, $g(\theta,t)$ and $d(t)$ known, so the control law can be given:

$$u^*(t) = g(\boldsymbol{\theta},t)^{-1}[-f(\boldsymbol{\theta},t) - d(t) + \ddot{r} - k_1\dot{e} - k_2 e] \tag{10}$$

When $f(\theta,t)$, $g(\theta,t)$ and $d(t)$ unknown, $u^*(t)$ can be approximated with fuzzy system.

$$u_{fz}(s,\boldsymbol{\alpha}) = \boldsymbol{\alpha}^{\mathrm{T}}\xi^{\mathrm{T}} \tag{11}$$

Where $\boldsymbol{\alpha} = [\alpha_1 a_2 \cdots \alpha_m]^{\mathrm{T}}, \boldsymbol{\xi} = [\xi_1 \xi_2 \cdots \xi_m]^{\mathrm{T}}$.

$$\xi_i = \frac{\omega_i}{\sum_{i=1}^{m} \omega_i} \tag{12}$$

With fuzzy approximation theory, can approach $u^*(t)$ with an optimal fuzzy system $u_{fz}(s,\hat{\boldsymbol{\alpha}})$. So:

$$u^*(t) = u_{fz}(s,\boldsymbol{\alpha}^*) + \varepsilon = \boldsymbol{\alpha}^{*\mathrm{T}} + \varepsilon \tag{13}$$

Where ε is approximate error and satisfied with $|\varepsilon| < E$.

$$u_{fz}(s,\hat{\boldsymbol{\alpha}}) = \hat{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{\xi} \tag{14}$$

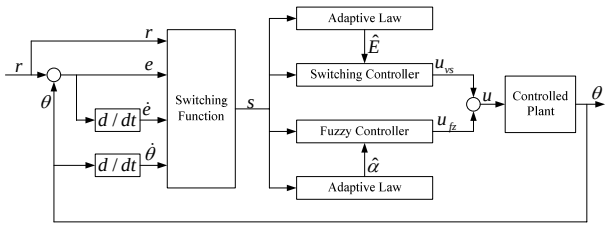


Figure 1. Adaptive Fuzzy Sliding Mode controller

Where $\hat{\boldsymbol{\alpha}}$ is the estimation of $\boldsymbol{\alpha}^*$.

Compensate the error between u^* and u_{fz} with switching control rule u_{vs} and the whole control law is:

$$u(t) = u_{fz} + u_{vs} \tag{15}$$

Adaptive control algorithm

The adaptive fuzzy sliding mode controller as shown in Fig. 1.

$$\tilde{u}_{fz} = \hat{u}_{fz} - u^* = \hat{u}_{fz} - u_{fz}^* - \varepsilon \tag{16}$$

Define $\tilde{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*$ and we can get:

$$\tilde{u}_{fz} = \tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{\xi} - \varepsilon \tag{17}$$

$$\dot{s}(t) = \ddot{e}(t) + k_1\dot{e}(t) + k_2 e(t) \tag{18}$$

$$u^*(t) = g(\boldsymbol{\theta},t)^{-1}[g(\boldsymbol{\theta},t)u(t) - \dot{s}(t)] \tag{19}$$

$$\dot{s}(t) = g(\boldsymbol{\theta},t)[u_{fz} + u_{vs} - u^*(t)] \tag{20}$$

Define Lyapunov function as:

$$V_1(t) = \frac{1}{2}s^2 + \frac{g(\boldsymbol{\theta},t)}{2\eta_1}\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\tilde{\boldsymbol{\alpha}} \tag{21}$$

$$\begin{aligned}\dot{V}_1(t) &= s(t)\dot{s}(t) + \frac{g(\boldsymbol{\theta},t)}{\eta_1}\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\dot{\tilde{\boldsymbol{\alpha}}} \\ &= g(\boldsymbol{\theta},t)\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\left(s(t)\boldsymbol{\xi} + \frac{1}{\eta_1}\dot{\tilde{\boldsymbol{\alpha}}}\right) \\ &\quad + s(t)g(\boldsymbol{\theta},t)(u_{vs} - \varepsilon)\end{aligned} \tag{22}$$

In order to realize $\dot{V}_1 \le 0$, the adaptive law and switching control can be given as (23), (24).

$$\dot{\tilde{\boldsymbol{\alpha}}} = \dot{\hat{\boldsymbol{\alpha}}} = -\eta_1 s(t)\boldsymbol{\xi} \tag{23}$$

$$u_{vs} = -E(t)\,\mathrm{sgn}(s(t)) \tag{24}$$

Then:

$$\dot{V}_1(t) = -E(t)|s(t)|\,g(\boldsymbol{\theta},t) - \varepsilon s(t)g(\boldsymbol{\theta},t) \le 0 \tag{25}$$

Replace $E(t)$ with $\hat{E}(t)$ and then:

$$u_{vs} = -\hat{E}(t)\,\mathrm{sgn}(s(t)) \tag{26}$$

Here define trace error:

$$\tilde{E}(t) = \hat{E}(t) - E(t) \tag{27}$$

$$\begin{aligned}V(t) &= V_1(t) + \frac{g(\boldsymbol{\theta},t)}{2\eta_2}\tilde{E}^2 \\ &= \frac{1}{2}s^2(t) + \frac{g(\boldsymbol{\theta},t)}{2\eta_1}\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\tilde{\boldsymbol{\alpha}} + \frac{g(\boldsymbol{\theta},t)}{2\eta_2}\tilde{E}^2\end{aligned} \tag{28}$$

$$\dot{V}(t) = \dot{V}_1(t) + \frac{g(\boldsymbol{\theta},t)}{\eta_2}\tilde{E}\dot{\tilde{E}} \tag{29}$$

Here we take adaptive Law with:

$$\dot{\hat{E}}(t) = \eta_2|s(t)| \tag{30}$$

And then we get:

$$\begin{aligned}\dot{V}(t) &= -\hat{E}(t)|s(t)|g(\boldsymbol{\theta},t) - \varepsilon s(t)g(\boldsymbol{\theta},t) \\ &\quad + (\hat{E}(t) - E)|s(t)|g(\boldsymbol{\theta},t) \\ &= -\varepsilon s(t)g(\boldsymbol{\theta},t) - E|s(t)|g(\boldsymbol{\theta},t) \\ &\le |\varepsilon||s(t)|g(\boldsymbol{\theta},t) - E|s(t)|g(\boldsymbol{\theta},t) \\ &= -(E - |\varepsilon|)|s(t)|g(\boldsymbol{\theta},t) \le 0\end{aligned} \tag{31}$$

III. APPLICATION IN INVERTED PENDULUM

Mathematical models of inverted pendulum

The dynamic equation of single inverted pendulum can be given [17]:

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = \dfrac{g\sin x_1 - mlx_2^2 \cos x_1 \sin x_1/(m_c+m)}{l(4/3 - m\cos^2 x_1/(m_c+m))} \\ \qquad + \dfrac{\cos x_1/(m_c+m)}{l(4/3 - m\cos^2 x_1/(m_c+m))}u(t) + d(t) \end{cases} \tag{32}$$

Where x_1 is angular position and x_2 is the velocity of the pole respectively, $g = 9.8\text{m/s}^2$, m_c is the mass of cart and $m_c = 1\text{kg}$, m is the mass of pole and $m = 0.1\text{kg}$,here $l = 0.5\text{m}$ is half lengthen of pole, u is control input, $d(t)$ is external disturbance and here are sine function $d(t) = 20\sin(2\pi t)$.

Simulation and conclusions

In order to prove the control ability of fuzzy sliding mode controller, we take simulation on linear and nonlinear system respectively; the linear system can be given as [18]:

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = -25x_2 + 133u(t) + d(t) \end{cases} \tag{33}$$

Here we take $d(t) = 20\sin(2\pi t)$, position tracking reference input was $r(t) = 0.2\sin(\pi t + \dfrac{\pi}{2})$, take the parameters $k_1 = 150, k_2 = 200$, so the sliding mode surface can be given as:

$$s(t) = \dot{\theta}(t) - \int_0^t [\ddot{r}(t) - 150\dot{e}(t) - 200e(t)]dt \tag{34}$$

The position tracking for sine function as shown in Fig.2, the control input as shown in Fig.4, the error as shown if Fig.5 and switching gain variation as shown in Fig.6.

Also simulation with nonlinear system of inverted pendulum above mentioned, here we take the parameters $k_1 = 10, k_2 = 25$, position tracking reference input was $r(t) = 0.2\sin(\pi t + \dfrac{\pi}{2})$,so the sliding mode surface can be given as:

$$s(t) = \dot{\theta}(t) - \int_0^t [\ddot{r}(t) - 10\dot{e}(t) - 25e(t)]dt \tag{35}$$

The initial state of single inverted pendulum is $\begin{bmatrix} \pi \\ 60 \\ 0 \end{bmatrix}$, simulated with fuzzy sliding mode control method that proposed in this paper, the position tracking for sine function as shown in Fig.2, we can see the tracking precision is high, the control input as shown in Fig.3, the error as shown in Fig.4 and switching gain variation as shown in Fig.5.

On the basis of the mathematical models of single inverted pendulum, FSMC controller can be applied into linear or nonlinear system successfully. The limitation of uncertainty bounds was released and the time derivative of the proposed sliding manifold was continuous, the chattering phenomenon in the sliding mode control was

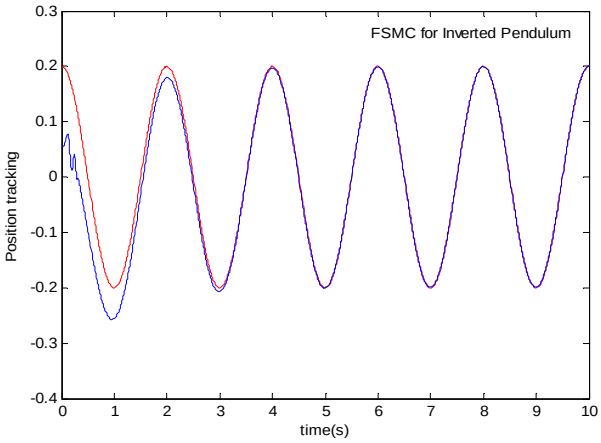


Figure 2. Posintion tracking for inverted pendulum

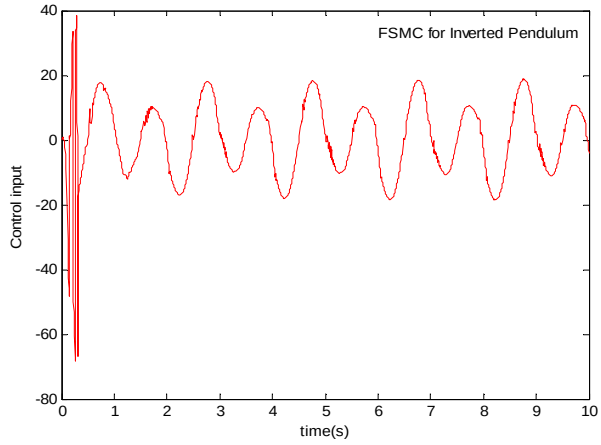


Figure 3. Control input for inverted pendulum

alleviated by fuzzy switching, the high robustness and precision performance of control performance obtained.

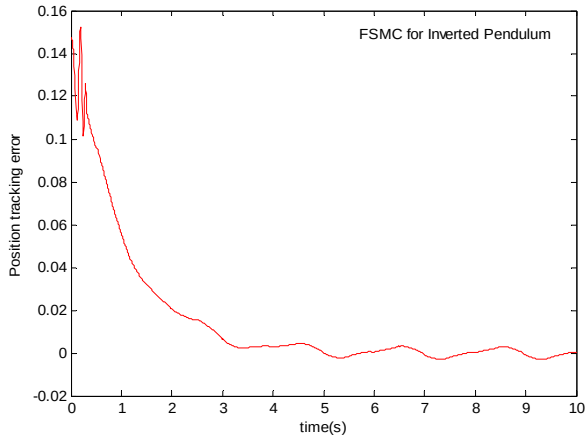


Figure 4. Position tracking error for inverted pendulum

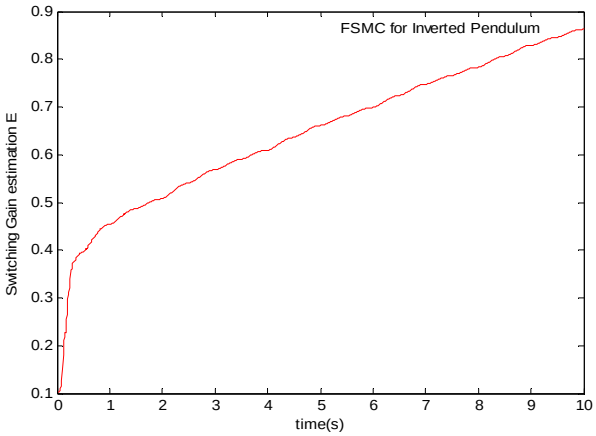


Figure 5. Switching gain for inverted pendulum

ACKNOWLEDGMENT

It is a project supported by natural science research in office of education, HeNan Province(2008A510014), Also the project was supported by XuChang University.

REFERENCES

[1] Sira R H, Llanes S O. Adaptive dynamical sliding mode control via backstepping. Proceedings of the 32nd IEEE conference on Decision and Control, 1993(2):pp.1422-1427.

[2] JIA Nuo, WANG Hui. Nonlinear Control of an Inverted Pendulum System based on Slinding mode method. ACTA Analysis Functionalis applicata, 2008,9(3): pp.234-237.

[3] Kawamura A, Itoh H, Sakamoto K. Chattering reduction of disturbance observer based sliding modecontrol. IEEE Transactions on Industry Applications, 1994,30(2), pp.456-461.

[4] Q. P. Ha, Q.H. Nguyen, D. C. Rye, H. F. Durrant-Whyte. Fuzzy sliding-mode controllers with applications. IEEE Transactions on Industry Applications, 2001,48(1), pp.38-41.

[5] Zhuang K Y, Su H Y, Chu J, Zhang K Q. Globally stable robust tracking of uncertain systems via fuzzy integral sliding mode control. Proceedings of the 3th World Congress on Intelligent control and Automation, 2000,pp.1824-1831.

[6] Ryu S H, Park J H. Auto-tuning of sliding mode control parameters using fuzzy logical. American Control Conference, 2001, pp.618-623.

[7] Kim S W. Lee J J. Design of a fuzzy controller with fuzzy sliding surface. Fuzzy Sets and Systems, 1995,71(3). pp.359-367.

[8] Yoo B, Ham W. Adaptive fuzzy sliding mode control of nonlinear system., IEEE Trans. On Fuzzy Systems, 1998,6(2). pp.315-321.

[9] Liang C Y. Su J P. A new approach to the design of a fuzzy sliding mode controller. Fuzzy Sets and Systems, 2003,139. pp.111-124.

[10] Chen J Y. Expert SMC-based fuzzy control with genetic algorithms. Jouranl of the Franklin Institute,. 1999,336:pp.589-610.

[11] Grossimon P G, Barbieri E, Drakunov S. Sliding mode control of an inverted pendulum. Proceedings of the Twenty-Eighth southeastern Symposimu on system Theory, 1996, pp.248-252.

[12] Chen Y P, Chang J L, Chu S R. PC-based sliding mode control applied to parallet type double inverted pendulum system, Mechatronics, 1999(9):pp.553-564.

[13] Lin F J, Shen P H, Hsu S P. Adaptive backstepping sliding mode control for linear induction motor drive. IEEE Proceeding Electrical Power Application,2002149(3),pp.181-194.

[14] Tan Y L, Chang J, Tan H L, Hu J. Integral backstepping control and experimental implementation for motion system. Prooceedings of the 2000 IEEE International Conference on Control Applications, Anchorage, Alaska, USA, pp.25-27.

[15] Kanayama Y, Kimura Y, Miyazaki F, et al. A stable tracking contrtol method for autonomous mobile robot. IEEE International Conference on Robotics and Automation,1990,pp.384-389.

[16] LIU J K,MATLAB Simulation for Sliding Mode Control,2005,10, pp.237-279.

[17] JinKun LIU, Fuchun SUN. Chattering free adaptive fuzzy terminal sliding mode control for second order nonlinear system, Journal of Control Theory and Applications, 2006, 4: pp.385-391.

[18] JinKun LIU, Fuchun SUN. A novel dynamic terminal sliding mode control of uncertain nonlinear systems. Journal of Control Theory and Applications, 2007,5(2):pp.189-193

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 235-238

The Effect of Efficient Models on Cryptography

Xiaodong Su
Harbin University of Commerce, Harbin, P.R.China
suxd001@yahoo.cn

***Abstract*—In recent years, much research has been devoted to the synthesis of link-level acknowledgements; however, few have analyzed the synthesis of checksums. After years of essential research into von Neumann machines, we show the analysis of symmetric encryption, which embodies the intuitive principles of cryptoanalysis. In order to fulfill this intent, we disconfirm that despite the fact that IPv6 and simulated annealing can interact to realize this objective, congestion control can be made interposable, amphibious, and perfect.**

***Index Terms*—semantic models, networking, cryptography**

I. Introduction

Recent advances in stochastic communication and constant-time modalities have paved the way for symmetric encryption. The notion that security experts agree with thin clients is mostly considered typical. On a similar note, after years of extensive research into forward-error correction [27], we argue the deployment of vacuum tubes, which embodies the unfortunate principles of software engineering. This might seem counterintuitive but is buffeted by related work in the field. To what extent can the memory bus be deployed to fulfill this purpose?

Predictably, the disadvantage of this type of method, however, is that forward-error correction can be made "fuzzy", stochastic, and robust. The basic tenet of this approach is the investigation of Markov models. It should be noted that Ova turns the random symmetries sledgehammer into a scalpel. The shortcoming of this type of solution, however, is that the much-touted flexible algorithm for the visualization of Moore's Law by Sun is optimal. however, constant-time modalities might not be the panacea that systems engineers expected. Obviously, Ova may be able to be enabled to synthesize the partition table. Such a claim at first glance seems counterintuitive but fell in line with our expectations.

Our focus in this paper is not on whether object-oriented languages and local-area networks [27] can interact to answer this issue, but rather on proposing an application for semantic algorithms (Ova). While conventional wisdom states that this obstacle is always overcame by the evaluation of the Internet, we believe that a different method is necessary. Two properties make this solution ideal: we allow digital-to-analog converters to control Bayesian methodologies without the simulation of 802.11b, and also our heuristic observes multimodal models. For example, many frameworks develop red-black trees. Despite the fact that similar frameworks enable XML, we surmount this problem without simulating flip-flop gates.

In this position paper, we make three main contributions. We construct an algorithm for the exploration of the World Wide Web (Ova), showing that the seminal signed algorithm for the refinement of B-trees [32] is optimal. Second, we concentrate our efforts on showing that the Turing machine and scatter/gather I/O are continuously incompatible [31]. We disprove that the seminal flexible algorithm for the improvement of DNS by John Hennessy [38] runs in (n) time [10].

The rest of this paper is organized as follows. We motivate the need for robots. To overcome this challenge, we use extensible modalities to demonstrate that congestion control can be made unstable, scalable, and encrypted. Continuing with this rationale, we place our work in context with the prior work in this area. Similarly, we demonstrate the refinement of checksums. Finally, we conclude.

II. Framework

The properties of Ova depend greatly on the assumptions inherent in our design; in this section, we outline those assumptions. Figure 1 diagrams an analysis of spreadsheets. This is a confirmed property of Ova. Similarly, we scripted a 6-month-long trace showing that our design is unfounded. This is a typical property of our methodology. We consider an application consisting of n robots. This is crucial to the success of our work. Consider the early design by Sasaki and Miller; our architecture is similar, but will actually solve this riddle [25]. The design for our application consists of four independent components: reliable configurations, amphibious models, massive multiplayer online role-playing games, and the confusing unification of massive multiplayer online role-playing games and DHTs. Even though information theorists entirely believe the exact opposite, our system depends on this property for correct behavior.

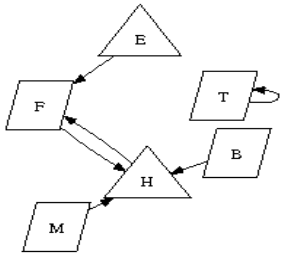


Figure 1: The relationship between our system and telephony

Suppose that there exists von Neumann machines such that we can easily synthesize the partition table. Consider the early framework by John McCarthy et al.;

our methodology is similar, but will actually solve this quandary. Thusly, the architecture that our framework uses is not feasible.

III. Implementation

In this section, we propose version 3.1.0, Service Pack 1 of Ova, the culmination of minutes of programming. The centralized logging facility contains about 723 semi-colons of SQL. of course, this is not always the case. Furthermore, it was necessary to cap the time since 1980 used by Ova to 106 percentile. Since our system provides collaborative communication, hacking the centralized logging facility was relatively straightforward. Overall, Ova adds only modest overhead and complexity to related flexible algorithms.

IV. Results

Our evaluation approach represents a valuable research contribution in and of itself. Our overall evaluation method seeks to prove three hypotheses: (1) that we can do little to affect a methodology's floppy disk throughput; (2) that work factor stayed constant across successive generations of NeXT Workstations; and finally (3) that effective signal-to-noise ratio stayed constant across successive generations of PDP 11s. note that we have intentionally neglected to synthesize RAM space. Unlike other authors, we have intentionally neglected to study energy [20,34,24,26,5]. We hope that this section proves Adi Shamir's emulation of Markov models in 1970.

A. Hardware and Software Configuration

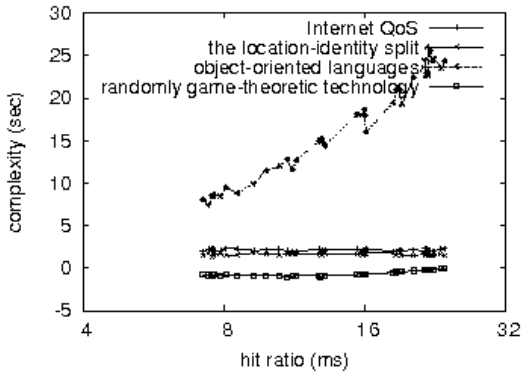


Figure 2: The 10th-percentile block size of our application, compared with the other systems

One must understand our network configuration to grasp the genesis of our results. We ran a robust prototype on DARPA's replicated cluster to quantify the collectively classical nature of extremely scalable algorithms. Had we deployed our network, as opposed to simulating it in courseware, we would have seen amplified results. Primarily, we reduced the effective flash-memory space of DARPA's desktop machines. This step flies in the face of conventional wisdom, but is essential to our results. Next, we doubled the seek time of our decommissioned UNIVACs. We added some tape drive space to our 1000-node overlay network. Along these same lines, we doubled the RAM speed of our constant-time testbed to discover the latency of our system. Along these same lines, we doubled the time since 1999 of our psychoacoustic testbed. Our objective here is to set the record straight. Lastly, we added some FPUs to UC Berkeley's desktop machines.

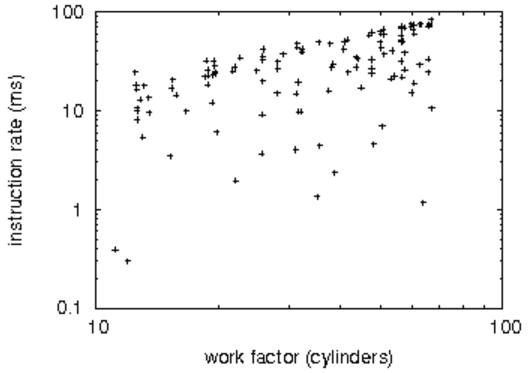


Figure 3: The effective work factor of Ova, compared with the other approaches

Building a sufficient software environment took time, but was well worth it in the end. We added support for Ova as a wired embedded application. All software was compiled using AT&T System V's compiler linked against empathic libraries for synthesizing online algorithms. Second, all software was hand assembled using Microsoft developer's studio with the help of Ken Thompson's libraries for opportunistically investigating parallel, computationally exhaustive systems. We made all of our software is available under a Sun Public license.

B. Dogfooding Ova

Given these trivial configurations, we achieved non-trivial results. We ran four novel experiments: (1) we ran active networks on 73 nodes spread throughout the millenium network, and compared them against 4 bit architectures running locally; (2) we measured instant messenger and RAID array throughput on our mobile telephones; (3) we ran object-oriented languages on 26 nodes spread throughout the sensor-net network, and compared them against I/O automata running locally; and (4) we asked (and answered) what would happen if opportunistically partitioned randomized algorithms were used instead of agents. All of these experiments completed without unusual heat dissipation or unusual heat dissipation.

Now for the climactic analysis of the second half of our experiments. While this result at first glance seems perverse, it has ample historical precedence. The key to Figure 2 is closing the feedback loop; Figure 3 shows how Ova's effective latency does not converge otherwise. Second, the many discontinuities in the graphs point to degraded sampling rate introduced with our hardware upgrades. Third, the key to Figure 2 is closing the feedback loop; Figure 3 shows how our application's block size does not converge otherwise.

Shown in Figure 3, experiments (3) and (4) enumerated above call attention to our framework's hit

ratio. The results come from only 0 trial runs, and were not reproducible [6]. Of course, all sensitive data was anonymized during our software simulation. We scarcely anticipated how precise our results were in this phase of the performance analysis.

Lastly, we discuss the first two experiments. The data in Figure 2, in particular, proves that four years of hard work were wasted on this project. Along these same lines, note that information retrieval systems have more jagged expected hit ratio curves than do modified compilers. The many discontinuities in the graphs point to muted interrupt rate introduced with our hardware upgrades.

V. Related Work

A major source of our inspiration is early work by John Hopcroft on erasure coding [14]. Furthermore, instead of developing "smart" symmetries, we surmount this quandary simply by controlling the analysis of massive multiplayer online role-playing games. Recent work by Sato [7] suggests a methodology for providing low-energy information, but does not offer an implementation [32]. Similarly, the original solution to this grand challenge by David Patterson was considered private; nevertheless, such a hypothesis did not completely overcome this obstacle. The little-known solution by John Hopcroft et al. [34] does not evaluate the analysis of superpages as well as our solution [7,18]. Thusly, if throughput is a concern, our algorithm has a clear advantage. Although we have nothing against the prior method by Deborah Estrin [21], we do not believe that method is applicable to algorithms.

A. Interposable Models

Several decentralized and pseudorandom methodologies have been proposed in the literature. We had our solution in mind before White and Raman published the recent much-touted work on the construction of journaling file systems [4,3]. Our design avoids this overhead. A litany of previous work supports our use of the simulation of active networks. Recent work by D. Lee et al. [31] suggests a methodology for requesting the refinement of the World Wide Web, but does not offer an implementation. Similarly, the choice of active networks in [30] differs from ours in that we synthesize only important models in Ova [12,40]. The only other noteworthy work in this area suffers from ill-conceived assumptions about the UNIVAC computer [16,23]. These heuristics typically require that interrupts can be made peer-to-peer, embedded, and event-driven [29,28,36], and we disproved in our research that this, indeed, is the case.

B. Perfect Archetypes

We now compare our approach to prior random methodologies solutions [13,5,35,19,33,7,22]. Furthermore, we had our approach in mind before Williams and Robinson published the recent seminal work on virtual communication [3,21,6]. Therefore, comparisons to this work are unfair. Although we have nothing against the prior approach by B. Wilson et al., we do not believe that approach is applicable to cryptography [8]. This work follows a long line of prior approaches, all of which have failed [1,17].

A number of prior methodologies have developed public-private key pairs, either for the improvement of architecture [39,9] or for the evaluation of Boolean logic [11]. The well-known framework by Anderson and Martin does not prevent the partition table as well as our method. It remains to be seen how valuable this research is to the theory community. Unlike many related methods, we do not attempt to analyze or locate the World Wide Web [2]. It remains to be seen how valuable this research is to the machine learning community. Instead of deploying the confirmed unification of the World Wide Web and spreadsheets [37], we fulfill this mission simply by studying psychoacoustic theory. This is arguably ill-conceived.

VI. Conclusion

To fulfill this objective for trainable epistemologies, we explored an analysis of suffix trees [15]. We probed how compilers can be applied to the improvement of hash tables. The construction of IPv4 is more intuitive than ever, and our methodology helps cryptographers do just that.

In conclusion, we showed in our research that 802.11b can be made lossless, distributed, and autonomous, and Ova is no exception to that rule. Continuing with this rationale, we also explored a scalable tool for refining multicast systems. Further, we concentrated our efforts on proving that suffix trees can be made distributed, homogeneous, and robust.

References

[1] C. Bhabha, "The influence of read-write methodologies on software engineering," Tech. Rep. 363-47-8325, Intel Research, Feb. 2004.

[2] P. Brown., and D. Qian, "A case for IPv7," *Journal of Amphibious*, Lossless Technology 77 ,1999, pp. 77-88.

[3] D. Clark, S. Hawking, and Q. Zhou, "Real-time epistemologies for the lookaside buffer," In Proceedings of PLDI, June 2002.

[4] J. Cocke, D. Knuth, R. Floyd, V. Maruyama, and T. Leary, "The relationship between robots and superblocks with Reek," In Proceedings of IPTPS, Jan. 1996.

[5] E. Codd, "A methodology for the understanding of object-oriented languages," In Proceedings of the Symposium on Interposable, Replicated Algorithms, Dec. 1999.

[6] P. ErdÖS, Y. Nehru, I. Bhabha, R. Hamming, V. Williams, and S. Floyd, "Decoupling active networks from consistent hashing in simulated annealing," Journal of Efficient, Replicated Information, Nov. 2001, pp. 57-67.

[7] D. Estrin, G. Watanabe, Z.H. Kobayashi, Z. Thomas, E. Schroedinger, R. Stearns, and D. Patterson, "Active networks considered harmful," In Proceedings of the Workshop on Stable, Embedded Configurations, Nov. 2005.

[8] H. Garcia-Molina, and H. Harris, "Psychoacoustic communication for the location-identity split," In Proceedings of the Workshop on Perfect, Secure Symmetries, June 2000.

[9] A. Gupta, D. Estrin, and A. Yao, "Deconstructing Voice-over-IP using VINE. In Proceedings of NDSS", May 1992.

[10] S. Gupta, “Highly-available, pervasive configurations,” In Proceedings of SOSP, June 2004.

[11] X. Haitao, W. Kahan, and B. White, “Understanding of RAID,” Journal of "Smart", Real-Time Epistemologies 23 ,Dec. 2004, pp. 52-65.

[12] C. Hoare, “Abele: A methodology for the improvement of information retrieval systems that paved the way for the simulation of the location- identity split,” OSR 40, Feb. 2004, pp. 1-13.

[13] C. Hoare, and X. Haitao, “Visualizing the World Wide Web and RAID,” In Proceedings of OOPSLA, June 2004.

[14] W. Ito, and M. Gayson, “On the visualization of massive multiplayer online role-playing games,” In Proceedings of the Workshop on Cacheable, Real-Time Information, Apr. 1990.

[15] T. Jackson, “A methodology for the emulation of the producer-consumer problem,” Journal of Relational Symmetries, vol. 9 ,Aug. 1999, pp. 154-193.

[16] X. Kobayashi, C. Leiserson, R. Morrison, and G. Anderson, “Investigating evolutionary programming and the Ethernet,” Journal of Modular, Psychoacoustic Configurations vol. 63 ,May 2004, pp.70-82.

[17] J. Kubiatowicz, and R. Martinez, “Contrasting IPv6 and IPv6,” In Proceedings of FPCA, Sept. 2005.

[18] K. Martin, V. Thomas, B. Martinez, R. Karp, D. Engelbart, and E. Schroedinger, “Urania: Improvement of evolutionary programming,” IEEE JSAC vol. 19, Dec. 2003, pp.42-51.

[19] J. McCarthy, M. Kaashoek, and R. Reddy, “Contrasting randomized algorithms and Internet QoS,” In Proceedings of the Workshop on Replicated, Perfect Configurations, Jan. 1999.

[20] K. Miller, and N. Chomsky, “Ubiquitous information for the memory bus,” In Proceedings of VLDB, June 2002.

[21] W. Miller, “Homogeneous information for context-free grammar,” In Proceedings of the Conference on Decentralized Algorithms, Apr. 2003.

[22] R. Milner, “Consistent hashing considered harmful,” In Proceedings of OOPSLA, Feb. 1992.

[23] I. Moore, “Improving DNS and Internet QoS,” Journal of Bayesian, Stable Communication vol.5 Aug. 2000, pp. 74-80.

[24] D. Ritchie, “Improving simulated annealing and active networks using DotedGuib,” Tech. Rep. 9229-9244, MIT CSAIL, Jan. 2003.

[25] E. Sasaki, M. Takahashi, and J. Hartmanis, “Amphibious, "smart" technology for rasterization,” In Proceedings of the Symposium on Ubiquitous, Homogeneous Models, Jan. 2000.

[26] R. Sato, and A. Turing, “Low-energy algorithms,” In Proceedings of the Workshop on Modular, Modular Models, July 2003.

[27] S. Shenker, and M. Kaashoek, “A case for link-level acknowledgements,” In Proceedings of the USENIX Security Conference, Oct. 1986.

[28] T. Sivaraman, and I. White, “Constructing interrupts using pseudorandom algorithms,” Journal of Symbiotic, Wireless Theory 79 ,May 2003, pp.83-100.

[29] J. Smith, “Investigating write-ahead logging and sensor networks,” In Proceedings of OSDI ,Sept. 1992.

[30] O. Smith, O. Sato, Z. Jones, A. Newell, and C. Hoare, “A case for the memory bus,” In Proceedings of FOCS, Apr. 1999.

[31] L. Subramanian, “Deconstructing the location-identity split,” Journal of Modular Methodologies 67, Dec. 1990, pp. 80-101.

[32] B. Sun, and A. Shastri, “Decoupling Smalltalk from Internet QoS in erasure coding,” OSR 44 ,Apr. 1999, pp. 71-88.

[33] A. Tanenbaum, “Lambda calculus considered harmful,” In Proceedings of MICRO, Nov. 1999.

[34] T. Taylor, A. Wang, A. Newell, M. Zhou, and Z. Watanabe, “Comparing reinforcement learning and replication,” Journal of Classical Information 1, Feb. 2003, pp. 1-11.

[35] D. Thomas, N. Watanabe, D. Patterson, Y. Gupta, M. Rabin, W. White, and I. Newton, “Refining web browsers using secure communication,” Journal of Scalable, Low-Energy Theory 63, June 1992, pp. 156-194.

[36] V. Williams, “Improving fiber-optic cables and Internet QoS,” In Proceedings of the Conference on Modular, Efficient Theory, Nov. 1998.

[37] P. Wu, “Sensor networks considered harmful,” In Proceedings of NDSS, July 2000.

[38] P. Wu, X. Haitao, and J. Wang, “Enabling symmetric encryption and the Internet using LITHIC,” In Proceedings of IPTPS, Feb. 2003.

[39] V. Wu, J. Jackson, A. Newell, and R. Floyd, “Extensible symmetries,” Journal of Self-Learning, Pervasive Symmetries 4, Feb. 1994, pp. 1-14.

[40] A. Yao, and R. Maruyama, “Deploying the transistor using introspective communication,” In Proceedings of the Symposium on Optimal, Probabilistic Symmetries, May 1993.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 239-242

Heterogeneous Information System Integration Based on HUB-SOA Model

Shaobo Li [1], Yao Hu [1], Qingsheng Xie [1], and Guanci Yang [2]
[1] Key Laboratory of Advanced Manufacturing Technology , Guizhou University , Guiyang, China
lishaobo@gzu.edu.cn,huyao520@hotmail.com,qsxie@gzcad.com
[2] Chinese Academy of Sciences , Chengdu Institute of Computer Applications , Chengdu, China
guanci_yang@163.com

***Abstract*—Considering the problems of heterogeneous informa-tion system integration, service-oriented component of HUB-SOA model is proposed based on investigating E-HUB integration model and SOA integrated structure. Further, the key technical of the basic working principle, interface model, web service, system component and schema mapping are researched. What's more, WEB service and its description of registration criterion for platform are defined, and service component integration architecture based on middleware is constructed, and the mapping model of heterogeneous information integration is proposed. Finally, the framework of HUB-SOA model with various heterogeneous systems is implemented.**

***Index Term*—HUB-SOA model; SOA; Service component; Heter-ogeneous information system; System integration**

I. INTRODUCTION

With the development of the internet, enterprise informa- tionization has entered a period of rapid development, large and small "information island" has been appeared in the enterprise because lack of overall planning when built all kinds of information systems. The heterogeneous information systems are unable to reengineering business process, unable to provide valuable decision information for leaders. The information system promotion of enterprise development at the beginning, but now it changed to constraints the development of enterprises.

Order to solve the problem of heterogeneous systems interconnection and communication of different protocols between the systems at the current. Appeared connection point to point integration model for solve one system with another system interconnection at the early, Hub-Spoke integration model which provide a application integration center[1], enterprise service bus integration model which connects multiple physical center-nodes, service-oriented integration model with service as a distributed component and so on[2]. For this integration models, Some of the code complexity , More difficult to maintain.some of lead the efficiency of the bottleneck easily,appear single point of failure problem[3]. So research integration model have a positive value and practical significance which based on service-oriented and service component.

II. HUB-SOA MODEL AND ITS INTERFACE MODEL

A. *Integration architecture of HUB-SOA model*

SOA (Service-Oriented Architecture) framework is an oriented architecture of service, it independent of the platforms of hardware, operating systems and programming languages[4]. E-HUB is a information platform which is based on open e-business standards and internet technology, Provide interconnection of information, data sharing and business interactive for enterprise by plug-in. SOA combined with Plug-type E-HUB[5], formation heterogeneous system integration framework adapt to enterprise. It includes two main aspects of Service-Oriented integration and service component integration which can "plug".

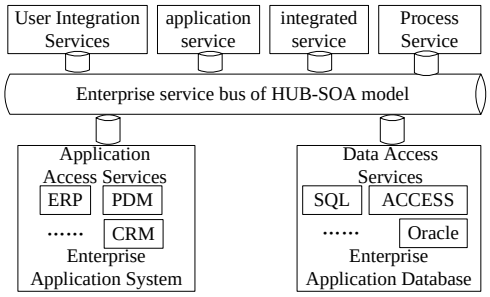


Figure 1. HUB-SOA model integration framework

HUB–SOA more open than the single Hub model, service component has unlimited expansion potential through service component; it truly embody everything is service, Hub deployment on the bus with the form of service, and constitutes the ESB of platform, it should be more flexible, this is the HUB-SOA. For traditional deficiency of SOA structure, We need a middle layer which can be achieved intelligent management of different services in SOA architecture, the framework shown in Fig. 1. HUB-SOA is a loosely coupled services and standards between of application integrate model which composition

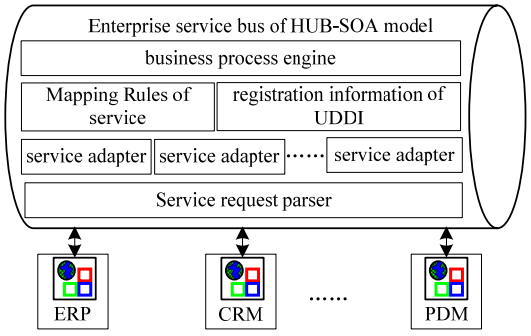


Figure 2. HUB-SOA interface model architecture

by service component, It can be used Service-Oriented Architecture, Component-oriented architecture, Message-oriented architecture, Event-driven architecture.

B. Interface model of HUB-SOA

This paper combination mode of E-HUB, the service adapter deployment in the service bus with the form of plug-in, publishing by Web Service, forming HUB-SOA integration architecture. As shown in Fig. 2.

Heterogeneous system service interface of HUB-SOA model is mainly composed by the following layers.

- Service request layer is the entrance of HUB-SOA, mainly including service parser and rule configurator. According to the mapping and configuration rules, operation and analysis service request object, generates operation sequences, get the data returned by service parser, and converts the data into system data by call service adapter.
- The data analysis layer is an important part of the HUB-SOA, and mainly composed of a plurality of service adapters. According to the rules described by platform data, it converts the dada into common formats which can be identified by platform, and submits the data to the business process engine.
- ESB implementation layer mainly includes business process engine, and is the core of the entire platform. It is in charge of the processing of business process for the overall platform, integrated data submitted from adapters and monitors the implementation of business.
- The service registry is mainly for realization the registration of the various heterogeneous systems service, equivalent to UDDI.
- Services mapping rules is mainly to complete the matching of mapping rules, and the choice of service adapters.

First, HUB-SOA integration framework takes component (adapter) as the internal services in ESB, Second, separating the service component from the ESB implementation engine, and convenient the "plug-in" for various adapters. By increase adapter components in architecture of SOA, the efficiency of the SOA is improved.

III. SERVICE COMPONENTS BASED ON HUB-SOA

Enterprise application integration is a tremendous and complex system, with components as its service units, and through the different service units composes the platform's ESB, and this way of constructing can fully support heterogeneous system. Through registering the service components on the platform, the HUB-SOA model achieves WEB data conversion fpr various heterogeneous systems and submit the datum to the next service component object. The platform can quickly achieve the enterprise business restructuring., through a variety of loose "plug-pull" service components.

The service component achieves Web service discovery, binding, calling etc, through parse for service information by components, in order to shield the details of the client calling back Web services, accesses to local or remote Web service transparently. Due to the differences data that WEB services received and returned provided by heterogeneous system, it needs convert the data which provided by each WEB services. The system structure and implementation principle of HUB-SOA model are as shown in Fig. 3.

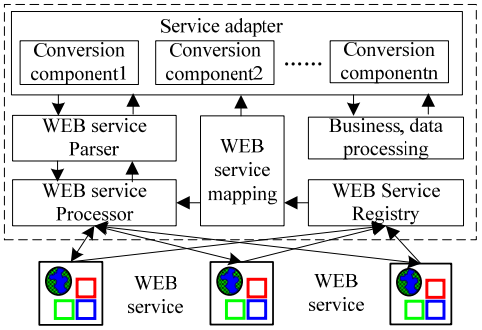


Figure 3. HUB-SOA Service Adapter Structure

IV. INTEGRATION AND MAPPING OF PATTERN ABOUT HETEROGENEOUS INFORMATION

In the heterogeneous information integration system, because data model of heterogeneous system is composed by different users, different time and location, it is independently designed based on different data models, may be exist a variety of differences and conflicts between them, in order to achieve the access and interactive of transparency to heterogeneous data, need to shield these difference and conflicts by research a way which can in the global level[6]. The mapped pattern found the inter-relationship between elements of pattern, then it set up the logical expressions of accord with semantics between elements of heterogeneous system through mapping. It is a key technologies about shield difference, conflicts and achieve integration of heterogeneous information system.

Definition 1. The mapping M of two data sources S1, S2 is a sets of mapping transformation model , has the following form.

M=(E,R,F)

E is the set of element in model, R is the set of relationships to elements, and F is the set of formulas by defined in the (S1, M), (M,M), (M, S2), thus R is called as map of S1, S2 based on M.

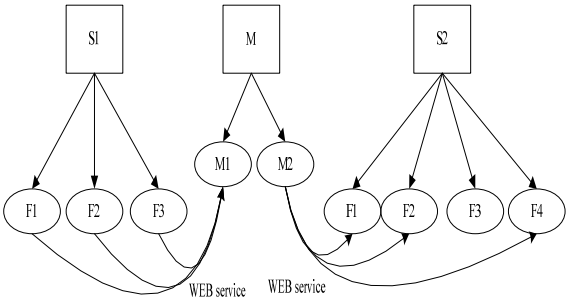


Figure 4. The mapped pattern of heterogeneous information

Element e of each mapping model M have one formulas which is defined in (S1, M), or (M, S2) at least, indicating the source or destination of elements. According to the definition of M, under the connections of mapping M, you can always find a path from a data source to another data source, so as to define and achieve data synchronization effectively. The model structure can be shown in Fig. 4.

Set S1 is the data source and S2 is the target data source. By Definition 1, generate data mapping elements, such as shown in Table 1.

TABLE I. THE ELEMENTS AND COMPUTING OF MAP M

Mapped element	Formula	Meaning
M	$S1 \overset{M}{=} S2$	Mapping by data
M1	$\ell(S1.F1) \overset{F}{=} M1, M1 \overset{M}{=} M2$, $M2 \overset{F}{=} S2.F1, S2.F2$	Data Source S1.-F1 achieve the d-ata synchronizati-on to S2.F1 and S2.F2 by mapping M
M2	$\ell(S1.F2, S1.F3) \overset{F}{=} M1$, $M1 \overset{M}{=} M2, M2 \overset{F}{=} S2.F4$	Data Source S1.F2 and S1.F3 achieve the data synchronization to S2.F4 by mapping M

- To a tuple in the data source S1, According to $\ell(S1.F1) \overset{F}{=} M1, M1 \overset{M}{=} M2, M2 \overset{F}{=} S2.F1, S2.F2$, ,element F1 generates mapping data by the service component F and mapping rules M, and modifies S2.F1 and S2.F2 into the mapping of $\ell(S1.F1)$,thus completes the transformation of elements.
- To a tuple in the data source S1, According to $\ell(S1.F2, S1.F3) \overset{F}{=} M1, M1 \overset{M}{=} M2, M2 \overset{F}{=} S2.F4$, realize the merger of data elements, and map them to S2.F4, thus complete the transformation of the second element.

Through the above manner and establishment of mapping rules, set up a standardized and unified mapping model from the data source to the target data, while the mapping requires a unified service component to achieve F and discernible service description specification.

V. DESIGN AND IMPLEMENTATION OF SYSTEM

A. System design

This article analysis the heterogeneous systems, designed a ESB model about web service of HUB-SOA, provides a way to solve these problems, the integration about Web Services of HUB-SOA is divided into the following sections.

1) Module of WEB Service, Register the server of WEB services, and set up the mapping between the WEB services.

2) Module of Enterprise business process, This module is mainly used for standardized business processes, and coordinating the data processing, build the service of bus enterprise.

3) Module of Data Conversion, According to mapping information, parse the data source data into the content of identify about target data and stored in the database.

4) Module of Data Management, This module achieve the functions of the data layer of the system, connected the mode of string through the control of the switch (local database or distance database), the string of local connection stored in the XML file, stored the information of mapping about database into the database.

This paper achieves data synchronization between heterogeneous systems, the steps design is as follows and data flow diagram shown in Fig. 5.

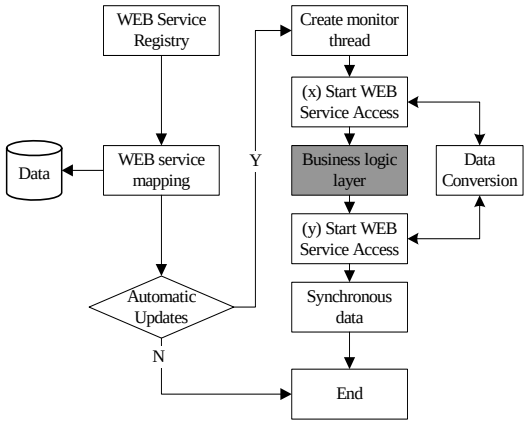


Figure 5. WEB Service Interface Data Flow Diagram

1) Registry of WEB Service, stored the address of the WEB service, as well as the namespace of data conversion, name of method into the database, when the service provider to return the data can be reflective of the method used to convert the database.

2) Mapped the service, set the startup mode of service and its start time of the service.

3) If the mapping is manual mode, then the end of the process, otherwise jumped to Step 4 operation;

4) The system will put the mapping information into the monitor thread;

5) Platform obtained the data of returned from the provider of service, call the corresponding data converter to convert the standard platform for the corresponding XML structure data.

6) Dealing with business processes according to the setting of ESB, and dealing with the data.

7) XML data will be replaced by URL parameters, and pass the parameters to the target service to achieve the synchronize of data.

B. system implementation

Web service registration is the basis of service interface, We need to record the address of WEB services, the method of analysis WEB services and namespace. The interface of running system as Fig. 6 and 7.

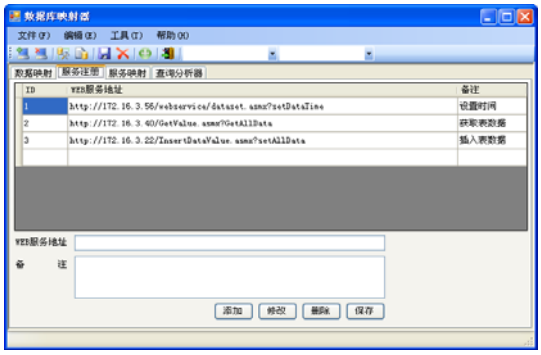


Figure 6. WEB services running interface(1)

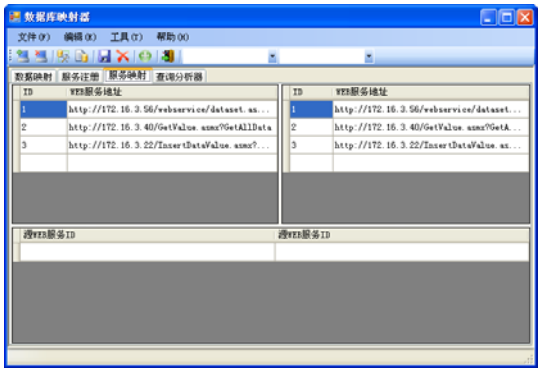


Figure 7. WEB services running interface(2)

VI. SUMMARIZE

Based on research of E-HUB and SOA framework, HUB-SOA model is proposed, Through further study the key technical of integrated information organization and description, Heterogeneous information integration and schema mapping, integration scheme and model based on middleware,Service-oriented interface model of heterogeneous systems;Through the study of xml language, description language specification based on xml of information integration is designed. the generality of integration model is improved, Business process that can really be portable and interoperable is realized. Through combination of these technical specifications, provides a viable, comprehensive and easily use framework for application of HUB-SOA architecture.

ACKNOWLEDGMENT

This research is supported by The National High Technology Research and Development Program of China(863 Program)(2006AA04Z130)and Outstanding Talents Project of Guizhou Province, Programs for Science and Technology Development of guizhou ([2008]3058, [2008]3047), Programs for Science and Technology Development of Guiyang.

REFERENCES

[1] Yuan Qingni, Xie Qingsheng, Li Shaobo, Research on Manufacturing Resources Collaboration Based on e-HUB Platform, Proceedings of the IEEE International Conference on Automation and Logistics, ICAL 2007, 2007, p 2566-2569.

[2] Meng Xiao-Jun, Zhang Xu, Ning Ru-Xin, Song, Yu, Enterprise integration platform framework based on Web services [J].Computer Integrated Manufacturing Systems, 2008, 14(5) : 891-897,961.

[3] Zhang Guang-Sheng, Jiang Cang-Jun, Tang Xian-Fei, Xu Yan, Specification and Verification of Service-Oriented Enterprise Application Integration System[J].Journal of Software, 2007,18(12): 3015-3030.

[4] Xu Li-Ming, Yao-Wen, Research and implementation of SOA developing framework[J].Application of Computers, 2008, 28(B06): 307-309.

[5] Zhou Gui-Xian, Xie Qing-Sheng, Hu Yao, E-LT integration to heterogeneous data information for SMEs networking based on E-HUB, ICNC 2008, v 5, p 212-216, 2008.

[6] Li Tao, Li Juan-Zi, Wang Ke-Hong, Heterogeneous Messages Match and Reuse in Web Services[J].Chinese Journal of Computers, 2006, 29(7) : 1038-1046.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 243-246

ITIL-based IT Service Management Applied in Telecom Business Operation and Maintenance System

Li Zhu[1], Meina Song[2], and Junde Song[2]
[1] School of Electrical Engineering in Beijing University of Posts and Telecommunications, Beijing, China
Email: zhulibupt @gmail.com
[2] School of Electrical Engineering in Beijing University of Posts and Telecommunications, Beijing, China
Email: {mnsong, jdsong}@butp.edu.cn

***Abstract*—Nowadays telecom operators have been aware of the significance of the telecommunications business Operation Management System, and have started to build an electronic operation and maintenance system, but how to plan the relationship between Operation Maintenance System and network management, as well as the function system which Operation Maintenance System want to achieve still lack of a unified model. IT Infrastructure Library (ITIL) is a internationally popular standard of IT operation and maintenance process. The article discussed how to introduce IT service management into the operation and maintenance of telecom business systems, it also designed management method which accommodate to telecom services operation and maintenance to improve the efficiency and quality of management to meet changing business needs and ensure systems' stable operation. The article focused on two design methods of management process which have stronger process uniformity and can effect a change more quickly, it also applied IT service management thinking to the actual maintenance process of telecom business systems.**

***Index Terms*—ITIL, ITSM, telecom business operation and maintenance.**

I. INTRODUCTION

Today, ITIL is recognized as the de-facto standard for managing enterprise IT and forms the basis of the international standard ISO 20000 [1]. ITIL was developed by the British Office of Government Commerce based on input from many industry leaders [2]. ITIL provides guidance and a common terminology for service management without being prescriptive about implementation. IT service management can enhance the management efficiency and quality to meet changing business needs. After improvement and development in 20 years, ITIL has evolved the best framework for IT service management. Based on ITIL and IT service management framework which is the core of processes and implementation, the article take a telecommunication company's IT service management processes for an example, studied logical framework, workflow and implementation of incident management and change management in IT service management system.

With the improvement in the level of information technology, the application environment of telecommunications enterprises has become more and more complicated, and customers are demanding an good quality of service [3]. Therefore, telecom operators need to introduce an ITSM (Information Technology Service Management) concept which is process-oriented and be with the core of customer satisfaction and service quality to make a comprehensive, focused, effective monitoring and management to their network, host and database systems. As a provider of Telecommunication business, in order to meet the requirements of telecom operators, has also been introducing various control processes actively, to improve their product quality of maintenance. Consequently, introducing IT Service Management to Telecommunication business Operation and Maintenance System is also important.

II. ITIL AND ITSM FOUNDATIONS

A. ITIL introductions

ITIL is currently the most widely-adopted approach to implement IT service management, and provides a set of best-practice guidelines for IT service management [4]. The ITIL guide breaks down the key principles of the IT service management discipline into the following sub-categories, which are collectively known as the ITIL

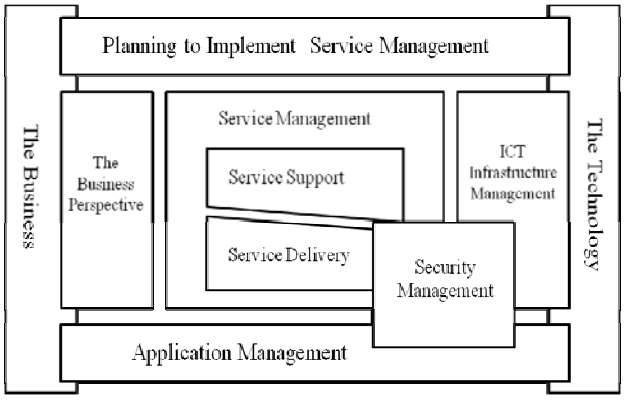


Figure 1. ITIL framework.

This work is supported by the National Key project of Scientific and Technical Supporting Programs of China under Grant No. 2008BAH24B04 and the innovation technology star program of Beijing under Grant No. 2007A045 and Program for New Century Excellent Talents in University under Grant No.NCET-08-0738.

framework, shown in Figure 1 [5][6].

The whole framework of ITIL includes six modules in the following [7]: planning to implement service management; the business perspective; service management; information and communication technology (ICT) infrastructure management; security management and application management. Service management, including service support and service delivery, is the core module of ITIL [8].

B. IT Service Management

ITSM based on ITIL, which integrates the best practices of global IT management and forms the normative truth standard to reduce effectively cost and improve the quality of service, is applied widely [9][10].

Service support and service delivery are considered to be at the heart of the ITIL framework for IT service management [11]. The Service Delivery module covers the processes involved in planning and delivering IT Services, and the Service Support module describes the processes required for those IT Services' daily support and maintenance [12].

IT Service Delivery function is closely related to the organization's annual planning cycle and assessment. Therefore, IT tactical management formed a strict functional logic. The five main functions are Service Level Management, IT Service Financial Management, IT Service Continuity Management, Capacity Management, and Availability Management [13].

IT Services Support focuses on the IT infrastructure's daily operations management, including five relevant basic management processes: Incident Management, Problem Management, Change Management, Configuration Management, and Release Management [13].

III. THE DESIGN OF THE INCIDENT MANAGEMENT PROCESS

An 'Incident' is any event which is not part of the standard operation of the service and which causes, or may cause, an interruption or a reduction of the quality of the service [14].

The objective of Incident Management is to restore normal operations as quickly as possible with the least possible impact on either the business or the user, at a cost-effective price [15].

A. Incident Management flowchart

Figure 2 is the UML activity chart which describes the process of incident management in maintenance of Telecommunication Business.

B. Role definition in Incident Management Process

Role definition in Incident Management Process is as follows:

1) Support: support is Audit and Manager of the Event Management Process, who is responsible for the examination of the event reports, to classify the incident to determine the lead Project Team of the incident, furthermore, to monitor the operation of the whole incident process.

2) Incident reporter: the general staff for Product

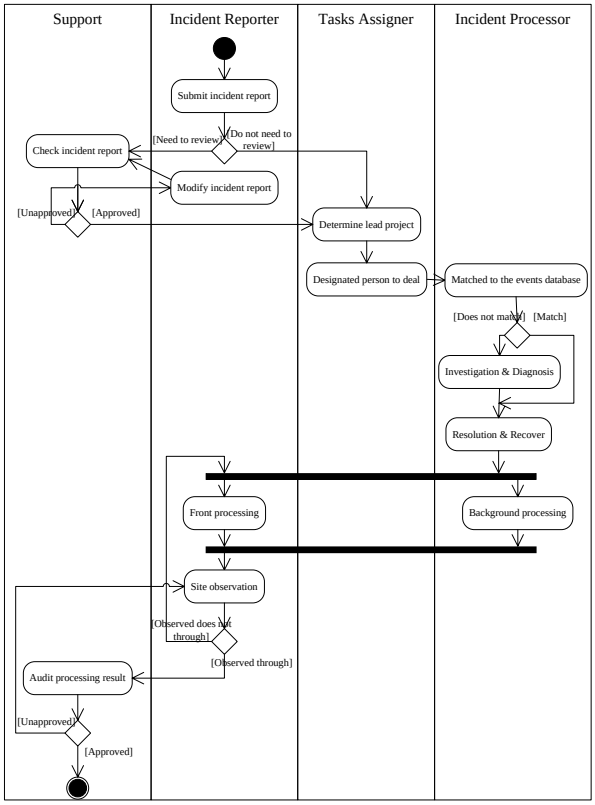


Figure 2. Incident management flowchart.

maintenance. After the incident, he needs to record Details of the process and Phenomenon when incident occur, then to report. And he is responsible for some event processing tasks.

3) The assigner of project: he is a project manager generally, responsible for the incident to take the lead, and assign specific personnel responsible for the handling of this incident.

4) Incident processor: The person who is responsible for the settlement of this incident as soon as possible. It can be added when necessary to tie in with other project team to solution the problem.

C. Incident Management Processes explain

Processing of incident is as follows:

1) Submit incident reports

Incident reporting is the entrance to the workflow. When event reporter is informed that the equipment or system events, he should first of all, implement the "submit incident reports" operation to create the incident case. According to the experience of workers at the scene is different, the incident report may be submitted directly to The assigner of the lead project, and may also be issued to support to audit, review and then sent to lead project.

2) Check incident reports

When support receive the report, he needs to check whether the format is correct, the events description is clear, the influence of events is described, event log is enough. If find the problem, support need to implement the "need to modify" operation, give report back to the

incident reporter; If there are no problems, support implement the "audited through the" operation, the incident report will be forwarded to the assigner of lead project.

3) Modify incident report

When Incident reports on events are give back from support, the reporter need to implement "re-submit" to submit report gain.

4) Determine lead project

Under normal, project manager also work as the assigner of project. The lead project of case should judge whether they are responsible for the case firstly after receiving incidence report from Engineering staff or support. If the analysis that the incident should be led by the project, and then implement "agreed to take the lead in" operation; if that should be led by the other projects, he needs to implement operation of "the request of other projects take the lead in".

5) Designated the person to deal with

The assigner of the lead project should considering the manpower of project and situation of the case when implement "agree to take lead", then specify appropriate person in charge and other projects to help.

6) Matching incidents

The person in charge of the case needs to verify the current incident and the incident in events library. If matched, he can resolve the event directing by the solution in incidents library. If there is no compatible incident, we should diagnosis and figure out the reason, and then resolve the incident.

7) Background processing

In the processing of incident, it is necessary for the processor and the reporter of incident to work together until the event recover.

8) Front processing

The reporter is responsible for implementation of solution provided by the processer, and check whether the solution is valid. if the solution is good and event recovers, the reporter can implement "resolve successful" operation, then the case entered the scene and observed stage; if that solution does not work, report implement "continue to analyze" operation, give it back to processer continue to resolve the incident.

9) Site observation

After the phenomenon in the incident is disappear, the needs of arrangements for a period of observation in according to the specific circumstances of the incident. If there are not anything unusual in the observation period, we can consider that the incident has now been ruled out, and then implement "observation pass" operation to close the case. If any problem is found by incident reporter in the observation period, then should implement "observation does not pass" operation and the case go back to the stage of "processing".

10) Audit processing result

Support checks the cases that have pass the observation. If found inadequate to deal with the results of verification, the project is not completely filled out, or there are other places do not meet the requirements, he should implement "verify does not pass" operation, then the case will back to observation status. Otherwise, should implement "verify pass" operation to close the case.

IV. The Design of the Change Management Process

Change management is the process during which the changes of a system are implemented in a controlled manner by following a pre-defined framework/model with, to some extent, reasonable modifications [16].

A. Change Management flowchart

Figure 3 is the UML activity chart which describes the process of change management in maintenance of Telecommunication Business.

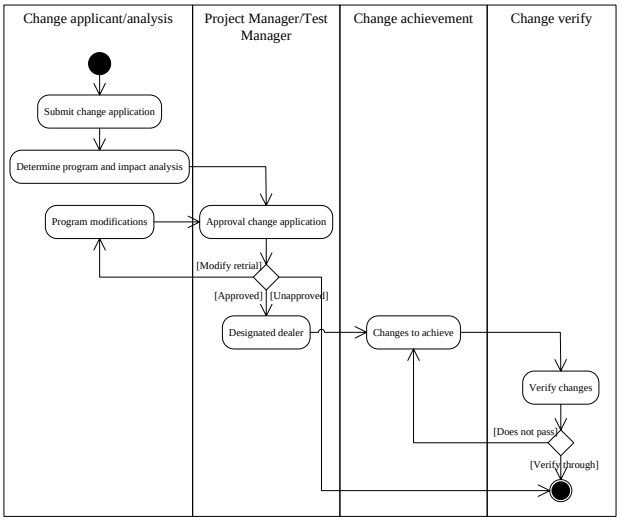


Figure 3. Change management flowchart.

B. Role definition in Change Management Process

Role definition in Change Management Process is as follows:

1) Change the applicant / analysis people: the general staff for the development, when the software bug was found or when the need to develop new features, submitted application to change.

2) Project Manager / Test Manager: General in charge of the project was responsible for approving the changes to the application, designating staff to deal with change.

3) The person who achieves change: the development of general staff for the project, responsible for the development of the realization of the change.

4) The person who verifies change: responsible to verify the effect of software changes, the general test for the project.

C. Incident Management Processes explain

Change management process details are as follows:

1) Submit an application to change

Submitting an application to change the software is the entrance of the workflow. Whether to repair program bug or develop new features, when the developers need to modify the software responsible for the maintenance of their own product, the applicant should first of implement

"submit an application to change" operation, to create case corresponding to the software change.

2) Determine program and impact analysis

After the submission of an application, we must accord to the change content to determine the basis of the initial solution and analysis impact based on the solution.

3) Approval change application

Project managers make a strategic decision based on the information Provide by applicant and analysis of change, considering the technical, market and other factors. The results divided into three types: approved changes, does not approve the changes, modify retrial.

4) Program modifications

If the result of examination is "modified retrial", analyst should modify the Program under the project manager's proposal, and then implement "modify complete" operation.

5) The officers designated somebody to deal with

The person who achieve Changes specified by the project manager on the implementation of "the approval of changes to" the operation.

6) Changes to achieve

Achieving changes in accordance with the program approved by the project manager, including code and documentation changes.

7) Verify changes

Verify changes through testing or other means (such as document review) to verify the results of changes. If the results meet the requirements, can implement "authentication through" operation to close software change case; otherwise, the implementation the operation of "authentication does not pass" and return to "change to achieve" continue to deal with.

8) Cancel changes

Project manager can implement "Cancel Change" operation at any time, suspend the ongoing changes in work, a direct closure of case. Project Manager should record the reasons for cancelling changes in the implementation of "Cancel Change" operation.

V. CONCLUSIONS

In this paper we've described an IT Service Management system which applied to maintenance of Telecommunication Business. We discussed two basic workflows of Service Management system: Incident Management and Change Management. Our experience shows that this system has regulated maintenance of Telecommunication Business providers' action to some extent. Furthermore, it also has enhanced the quality of maintenance as well as has opened up a way to improve customers' satisfaction.

ACKNOWLEDGMENT

The authors wish to thank Xiaoqi Zhang, Hongbo Guo, and Shu Zhang. This work was supported in part by a grant from PCN&CAD Center, BUPT.

REFERENCES

[1] ITIL The Key to Managing IT Services, Service Support Version 2.3, (TSO for OGC), 2000.
[2] Information technology Service management, (ISO/IEC 20000).
[3] P. Weill, IT Governance, "How Top Performers Manage IT Decision Rights for Superior Results," in Beijing, The Commercial Press, 2005.9.
[4] Doughty, K. (2003), "A Framework for Incident and Change Management," Enterprise Operations Management.
[5] "Enhance IT Infrastructure Library service management capabilities," http://www.ibm.com/developerworks/autonomic/library/ws itil/index.html.
[6] Bon J V, "IT Service Management and Introduction," in Canada, Van Haren Publishing, 2002.
[7] Office of Government Commerce (2000), IT lnfrastiucture Library, The Stationary Office, in London.
[8] Wardale, Dorothy. "4 Components of the module", Retrieved on 30 January 2009.
[9] ITIL Organization Structure, CEC Europe Briefing Papers, Version2.0, July 2002.
[10] Cao Hanping, Wang Qiang, Jia Suling, "Modern IT Service Management," Tsinghua University Press, 2005.
[11] Van Bon. J. (2002), "The Guide to IT Service Management," Addison, Wesley, in London.
[12] ISO/IEC, ISO 20000, "Information technology — Service management", (ISO) and (IEC), 2005.
[13] Jan Van Bon, "Foundations of IT Service Management," based on ITIL[M].Van Haren Publishing, 2005.
[14] Information technology Service management, (ISO/IEC 20000).
[15] ITIL Incident Management - The ITIL Open Guide.
[16] ITIL - The Key to Managing IT Services, Service Support Version 2.3, (TSO for OGC), 2000.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 247-249

Mine Cross-Level Location Sequences in RFID System

Kongfa Hu[1,2], Youwei Ding[1] , Ling Chen[1] , and Aibo Song[2]
[1] Department of Computer Science and Engineering, Yangzhou University , Yangzhou 225009, China
[2] School of Computer Science & Engineering, Southeast University, Nanjing 210096, China
Email: kfhu05@126.com

***Abstract*—With RFID technology being applied in more and more applications nowadays, many researchers have focused on how to manage and analyze RFID data efficiently. Obviously, location information, which is an important part of RFID data, has the characteristic of multiple conceptual levels. Existing methods of RFID data management mainly deal with the cases of non-hierarchy (single hierarchy) or multiple hierarchies. In this paper, we propose an efficient algorithm to mine the cross-level location sequences. It can find the location sequences with items at different conceptual levels to help different analysts making decisions. Experimental results have shown that our new algorithm is practical and effective.**

***Index Terms*—RFID(Radio Frequency Identification); cross-level; location sequence**

I. INTRODUCTION

RFID (Radio Frequency Identification) technology can automatically identify objects without line of sight. RFID reader sends radio-frequency waves to identify all the tags stuck on the moving products in its covered area, and each tag has a global unique EPC (Electronic Product Code). In a RFID system, for example a logistics network, some readers are deployed in different locations to collect the whole moving path of every product during the process of producing, transporting and selling. Since location information is an important part of moving paths, efficiently analysis of location information has a great influence on decisions making for analysts. Considering the complexity of the a RFID system, users only interest in the part of moving paths they are responsible for, such as the manager of a supermarket may only focus on products moving from warehouse to shelves and finally to checkouts. Therefore, how to mine cross-level location information from the moving paths becomes a hot research issue.

Previous research on RFID data management takes the conceptual hierarchies into account in the process of data storing and data cleaning. And in recent years, many methods has been proposed to mine RFID data [1,2,3], in which [3] presented a bitmap-based algorithm to efficiently mine multi-level location sequences. Unfortunately, there have not any efficient algorithm of mining cross-level location sequences in RFID datasets.

Although many methods of mining cross-level frequent itemsets had been presented [4], only a few researchers pay attention to the issue of cross-level sequential patterns mining because of the characteristics of sequences. Existing algorithms of cross-level sequential patterns mining mainly based on Apriori [5] theory and prefix projection [6], the former of which could generate numerous candidates when the number of frequent items is large, while the other of which would lead to a heavy I/O cost.

In this paper, we propose a bitmap-based algorithm, CL-LSM, to mine cross-level location sequences. Similar to ML-LSM [3], we focus on reducing I/O cost by storing the compressed representation of raw dataset into main memory and no extra scanning of raw data during mining process. But in CL-LSM, Apriori theory can be adopted, i.e. if a node is infrequent then its descendants at lower level are infrequent either. In addition, we introduce a pruning strategy to prune those sequences which are of no senseless in applications.

II. PRELIMINARIES AND RELATED WORK

The form of raw data RFID readers collect is <EPC, location, time>, and if an object stays a period of time at a location, we often store the compress record <EPC, location, time_in, time_out> of the series of raw data, where time_in and time_out are the time the tagged object came in and out of the location. Without considering the detailed time, we can integrate the tuples with the same EPC according to their generated time into one tuple named path sequence, which is in the form of $<(l_1,d_1),(l_2,d_2),\ldots,(l_n,d_n)>$, where (l_1,d_1) represents the object staying d_i units of time at location l_i , and d_i=*time_out - time_in*.

Location sequence is part of path sequence, which is in the form of $<l_1,l_2,..,l_n>$ with the means of an object moving from l_1 to l_2, then to ..., and finally to l_n. Therefore, the path database PDB should be transformed into location sequence database LDB, in which each tuple with the form of (lid,$<l_1,l_2,\ldots,l_m>$: w), where is lid the identifier of a location sequence $<l_1,l_2,\ldots,l_m>$, and w is the weight of the location sequence, i.e. the counts the location sequence appears in PDB. Each location is coded into a series of positive integers as introduced in [3].

Definition 1(cross low/high level) The levels the mining performed on are denoted as interest levels. The

lowest interest level is called cross low level, and the lowest interest level is called cross high level.

Let location item I=<$a_m a_{m-1}\ldots a_2 a_1$>, where a_i is the code at the i-th level, the aggregation (ancestor) of the k-th level of I is denoted as I(k)=<$a_m\ldots a_k b_{k-1}\ldots b_1$>, where 1<k<m, and b_i=* for 1<=i<=k-1.

Definition 2(cross frequent item) Each item at interest levels whose support is not less than the support threshold is called cross frequent items.

Definition 3 (cross-level location sequence) Location sequence *ls*=<$I_1,I_2,\ldots,I_m$>, where $I_i = < a_m^i a_{m-1}^i \cdots a_1^i >$ is coded location item, is a cross location sequence, if there exists an integer 1<=*h*<*m*, and $a_h^i \neq *$ and $a_{h-1}^i = *$ for *i*=1…*m*. If the support of *ls* is not smaller than the given support threshold, we call *ls* a frequent cross-level location sequence.

Our task is to mine all the frequent cross-level location sequences at interest levels in LDB.

III. MINE CROSS-LEVEL LOCATION SEQUENCES

The process of mining cross-level location sequences consists of three steps. First, we should find all the cross-level frequent items. Second optimize the location sequences database, and then get the bitmap representation of optimized database. Finally, mine all the frequent cross-level location sequences.

We can obtain the set of cross-level frequent items by scanning LDB once. Then we optimize the LDB according to the cross-level frequent items. We should first delete the obviously infrequent items, the items infrequent at the cross high level; as to the other items, we get their first frequent aggregations from cross low level to cross high level to replace the raw item. In the new sequence, if two consecutive items are the same, preserve only one. Finally, we get the bitmap representation of optimized database, i.e. the appearing sequence table, the bit sequence table and weight table, which is in the same spirit of [3]. The procedure of optimization is shown in figure 1.

```
Procedure Optimization
1    for i=1; i<=n; i++
2      L'=0;
3      for j=1; j<=Li; j++
4        for l=cll; l<=chl; l++
5          if Iij(l) ∈CFI then
6            L'++;
7            IiL'=Iij(l);
8            if L'>1 and IiL'=IiL'-1 then
9              L'--;
10           break;
```

Figure 1. Procedure of optimization

In this procedure, cll and chl are the cross low level and cross high level, n is the number of tuples in LDB, Li is the length of the i-th tuple in LDB, and L' denotes the length of optimized tuple, which reduces the memory storage of the bitmap representation. Note that, we only generate the bitmap representation of optimized database, instead of a new database stored on disk.

For example, given LDB in table 1, and the support threshold is 0.5. The levels between level 3 (raw data level), i.e. cross low level, and level 1, the cross high level, are interest levels. We can obtain the set of cross-level frequent items {211:8, 111:7, 122:6, 221:6, 311:6, 21*:11, 31*:11, 12*:9, 11*:8, 22*:6, 1**:11, 2**:11, 3**:11} and the optimized LDB as shown in table 2. From table 2, we can see that item "412" is deleted because "412", "41*" and "4**" are infrequent, and "212" is replaced with "21*" as "21*" is the first frequent aggregation of item "212" of interest levels from cross low level to cross high level.

In the process of mining location sequences, an item I and its ancestors of level k I(k), where cll<=k<=chl, cannot appear in a location sequence simultaneously. For example, a sequence means "a computer is moving from Beijing to China" is of no sense. Given a frequent cross-level location sequence s with length |s|=l, and the cross-level frequent items CFI, our task is to find all the frequent cross-level location sequences whose length is l+1 and prefix is s.

TABLE 1. RAW LOCATION SEQUENCES DATABASE

LID	LS	Weight
01	<111,211,212,412,312,321>	2
02	<111,122,212,221,311,331>	1
03	<111,122,212,221,311>	2
04	<112,121,411,211,312,331>	1
05	<111,112,121,211,312>	2
06	<122,211,221,311,321,421>	3

TABLE 2. OPTIMIZED LOCATION SEQUENCES DATABASE

LID	LS	Weight
01	<111,211,21*,31*,3**>	2
02	<111,122,21*,221,311,3**>	1
03	<111,122,21*,221,311>	2
04	<11*,12*,211,31*,3**>	1
05	<111,11*,12*,211,31*>	2
06	<122,211,221,311,3**>	3

First, we should find the cross-level frequent items that can appear in a same sequence with s, which is denoted as C(s). For each item *I* in C(s), extend s to s' by I-step introduced in [7,8], then compute the $FP_{s'}$ according to the current table of s and I, and finally get the count of s'. If support(s') is no less than the given support threshold, we return s' as a frequent cross-level location sequence.

IV. PERFORMANCE STUDY

To evaluate the performance of our proposed algorithm, we design a simple benchmark which is based on DC-CROSS-FMSM introduced in [6]. Our proposed algorithm is briefly written as CL-LSM (Cross-Level Location Sequences Mining). We conduct all the tests on a P4 2.8GHz PC with 2G memory, running Microsoft Windows XP Professional. The algorithms were implemented with Microsoft Visual C++ 6.0. With respect to data set generation, we employ the idea similar

in spirit to IBM generator. The convention for the data sets is as follows: L3C10I10T100K means the dimension (location dimension) contains 3 levels, the node fan-out factor (cardinality) is 10 (i.e. 10 children per node), the average length of a sequence is 10, and there are in total 100K tuples (path sequences) at the bottom level.

Here we only compare the running time of CL-LSM and benchmark when the tuples of dataset increasing from 100K to 500K. In Fig.2, we can see the comparison using parameters L=3, C=10, I=10 and the threshold of support is 0.5. It is clearly shown that the relative improvement increases with larger datasets, since bitmap-based algorithm costs little I/O cost.

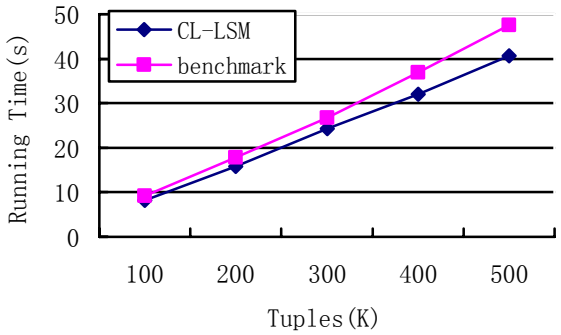


Figure 2. Running time over different datasets

V. CONCLUSIONS

In RFID application, different users have interest in different part of the whole moving path, which makes analysts find cross-level sequences to support decision making. Unfortunately, previous research typically focused on single level and multi-level sequences mining. In this paper, we propose an efficient algorithm CL-LSM to mine cross-level location sequences in RFID database. It is an extension of traditional bitmap-based sequential pattern mining algorithm.

CL-LSM is also a space-wasted method, because of the compacted bitmap representation.

VI. ACKNOWLEDGEMENTS

The research in the paper is supported by the National Natural Science Foundation of China under Grant No. 60773103 and 60673060;the Natural Science Foundation of Jiangsu Province under Grant No. BK2009697 and BK2008206; the "Six Talent Peaks Program" of Jiangsu Province of China; the Natural Science Foundation of Education Department of Jiangsu Province under Grant No. 08KJB520012 ;the 'Qing Lan' Project Foundation of Jiangsu Province of China.

REFERENCES

[1] H. Gonzalez, J. Han, X. Li, and D. Klabjan, "Warehousing and analyzing massive RFID data sets," Proceedings of the International Conference on Data Engineering, pp.83-87,2006.

[2] Z. Chen, G. Hong, K. Hu, and L. Chen, "Divide: Mining Closed Frequent Path for Commodities in Supply Chain," IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application, pp.944-948, 2008.

[3] Y. Ding, J. Yang, K. Hu and L. Chen, "Multi-level Frequent Location Sequences Mining in RFID Database," International Workshop on Computer Science and Engineering ,pp. 458-461,2009.

[4] T. Eavis and X. Zheng," Multi-level Frequent Pattern Mining," Proceedings of the 14th International Conference on Database Systems for Advanced Applications, pp.369 – 383, 2009.

[5] Y. L. Chen, and T. C. K. Huang, "A novel knowledge discovering model for mining fuzzy multi-level sequential patterns in sequence databases," *Data and Knowledge Engineering*, vol. 66,2008, pp.349-367.

[6] T. C. K. Huang, "Developing an efficient knowledge discovering model for mining fuzzy multi-level sequential patterns in sequence databases," Proceedings of the International Conference on New Trends in Information and Service Science, pp.362-371, 2009.

[7] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu, "Sequential Pattern Mining Using A Bitmap Representation," Proceedings of the International Conference on Knowledge Discovery and Data Mining , pp.429-435,2002.

[8] C. L. Wu, J. L. Koh, and P. Y. An, "Improved sequential pattern mining using an extended bitmap representation," Proceedings of the 16th International Conference of Database and Expert Systems Applications, pp.776-785, 2005

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 250-253

Optimal Model of Web Caching and Prefetching

Lei Shi [1], Yan Zhang [2], and Wei Lin [3]
[1] School of Information Engineering.Zhengzhou University,Zhengzhou,China
Email: shilei@zzu.edu.cn
[2] Henan PolyTechnic College,Zhengzhou,China
Email: iezhangy@yahoo.com.cn
[3]School of Software, Zhengzhou University
weilin@zzu.edu.cn

***Abstract*—Caching and prefetching play important roles in improving the quality of data access. Replacement and prefetching algorithm optimization is the core of caching and prefetching model research. First, Independent Reference Model and Markov Reference Model are analyzed and compared in this paper. And so as the Markov-based Prefetching Model. Then, based on the measurement of Relative Popularity and Byte Cost, optimal Web caching and prefetching model named PR PPM are presented and analyzed. Simulations show that the performance of optimal model.**

***Index Terms*—caching model; Web prefetchig model; Relative popularity; optimal model PR PPM**

I. INTRODUCTION

Caching and prefetching as effective approaches to explosive growth in Network users and Web service, and has been widely used in Web Proxy,P2P,Grid Computing and Wireless network. Bringing some of more popular items closer to end-users can improve the network performance and, therefore, reduce the download latency and network congestion. Web caching and prefetching are based on temporal locality of user sequence.

Independent Reference Model (IRM) and Markov Reference Model (MRM) are mostly used for Web caching Model at present. While Markov-based Prefetching Model is mostly used for prefetching. The design of replacement policy is always based on characteristic of request sequences. Therefore, to modeling on user request sequences and Web objects properties exactly and simply is so important, and we hope to find optimal policies under these factors to be pursued in systematic manner. This paper firstly analyzes and compares Web caching and prefetching models that are used nowadays, and then based on the measurement of Relative Popularity and Byte Cost, it presents an optimal Web caching and prefetching model PR PPM that satisfy different performance metrics.

The rest of this paper is organized as follows. Section 2 discusses the related work. In section 3, optimal Web caching and prefetching model are presented and analyzed. Simulations results are also analyzed in section 3. Section 4 contains the summary and conclusions.

Supported by the National Natural Science Foundation of China under Grant No.60472044

AP-PROC-CS-09CN005

II. RELATED WORK

Numerous studies show that [1][2][5] there are some regulations in Web environment: Web traces exhibit excellent temporal locality and spatial locality; Web object size follows a heavy-tailed distribution; The popularities of Web documents usually follow generalized Zipf's law distribution; Viewing the Web surfing as a random walk and the probability distribution of surfing depth follows a two-parameter inverse Gaussian distribution. The notations and their presentations used in this paper are illuminated in table 1.

A. Conventional cache model

Independent Reference Model (IRM): In the context of conventional caching techniques, the underlying working assumption is Independent Reference Model (IRM) [2][4]. Under IRM the miss rate (respectively, the hit rate) is minimized (respectively, maximized) by the policy A according to which a document is evicted from the cache if it has the smallest probability of occurrence

TABLE I.
THE NOTATIONS AND THEIR PRESENTATIONS

Notations	Presentations
N	Objects available in Web server
R	User request sequences
R_t	The t^{th} request in cache
S	Objects in cache
S_t	Objects in cache before R_t arrived
X	Cache states volume
X_t	Cache state at time t
V	Operation volume
V_t	Objects evicted from cache if R_t is not in cache
C	Restriction
A	Replacement algorithm
B	Cache size
s_i	The size of i
c_i	Replaced cost if i is not in cache
l_i	Download latency of i
P_i	Popularity of i
RP_i	Relatively popularity of i
U_i	Byte cost of i

(respectively, is the least popular) among the documents in the cache. IRM can be formalized as a multi-tuple, M= (B, V, A, R, C), thereinto:

(1) if $R_t \notin S_t$, $V_t \in S_t$; else $V_t \in \phi$.

(2) R is assumed to form i.i.d.

(3) A is mandatory.

(4) restriction C: $\forall$ i and j, $s_i=s_j$, $c_i=c_j$.

IRM doesn't consider the statistical information contained in the stream of requests and fails to capture temporary locality. However, it is simple enough to find effective policy.

Markov Reference Model (MRM): Another reference model often encountered in caching applications is the Markov Reference Model (MRM) [3][4] according to which requests are modeled by a stationary and ergodic Markov chain. MRM can also be formalized as a multi-tuple, M= (B, V, A, R, C), thereinto:

(1) if $R_t \notin S_t$, $V_t \in S_t$; else $V_t \in \phi$.

(2) R is prescribed by a Markov chain.

(3) Restriction C: Markov.

(4) The MRM specializes to the IRM.

The key property of MRM is Markov.MRM is good for independent distribution and request streams with less contact.

B. Exisiting prefetching model

A Markov model is a finite-state machine where the next state depends only on the current state. Associated with each arc of the finite-state machine network is the probability of making the given transition. When applied to the prediction of user accesses, each state represents the context of the user.This model can also be formalized as a multi-tuple, PM= (X, P, O, C), thereinto:

(1) X denotes the states space.

(2) P_{ij} is the transition probability from x_i to x_j.

(3) O is the length of context.

(4) Other restrains like threshold, pruning and so on.

The mostly used PPM Models are Standard PPM, LRS PPM Model and PB PPM Model. The comparisons of these typical PPM Models are given in table 2.

C. Web caching model

A total of N distinct cacheable objects {1, 2…, N} are available over all servers. For each t= 0, 1, 2…, n, the n-value rv R_t represents the t^{th} request presented at the cache. The stream of successive requests arriving at the cache is then captured by the sequence of rvs R={R_t,t=0,1,2,…,n}. The popularity of requests in the sequence {R_t,t=0,1,2,…,n} is defined as the pmf P, P_i is the popularity of document i.

We assume that the cache can contain at most M objects with $M \le N$ .We define the variable X_t as the state of the cache at time t=0, 1, 2…, n. The evolution of the cache is tracked by the collection X= {X_t, t=0, 1, 2…, n}, and is affected by the stream of incoming requests R, and by the policy A that produces the eviction actions V_t.

Based on the content of the cache evolves after the request R_t is handled , we have

$$S_{t+1}=\begin{cases} S_t & if\ R_t \in S_t \\ S_t + R_t & if\ R_t \notin S_t, |S_t| < M \\ S_t + R_t - V_t & if\ R_t \notin S_t, |S_t| = M \end{cases} \quad (1)$$

where $|S_t|$ denotes the cardinality of the set S_t, and $S_t + R_t - V_t$ is a subset of N obtained from S_t by adding R_t and removing V_t. The eviction action V_t at time t= 0, 1, 2…, n is dictated by a cache replacement policy. Variety of cache state is decided by:

(1) Cache states collection, X= {X_t, t=0, 1…};

(2) User request sequences, R= {R_t, t=0, 1…};

(3) Eviction actions produced by policy A, V= {V_t , t=0, 1…}.

We select X_t as the pair (S_t, R_t) for time t, therefore we have X_{t+1} = (X_t ,V_t, R_{t+1}) at time t+1.

Web caching model can be defined as a multi-tuple, M= (B，V，A，R，C)，thereinto:

(1) if $R_t \notin S_t$, $V_t \in S_t + R_t$, else $V_t \in \phi$;

(2) R forms Zipf-like distribution;

(3) A is randomized;

(4) Restriction C: whenever $i \neq j$, $s_i \neq s_j$, $c_i \neq c_j$.

Different from conventional caching model, under Web caching model, the objects have non-uniform costs (as we assimilate cost to size and variable retrieval latency), there exist correlations in the request streams and request streams form Zipf-like distribution, and object placement and replacement are optional upon a cache-miss.

TABLE II.
THE COMPARISONS OF TYPICAL PPM MODELS

Model	Based on	Adaptability	Space complexity
Standard PPM	Markov chain	no	high
LRS PPM	Longest Repeating Sequences	no	low
PB PPM	Popularity	no	low

II. OPTIMAL WEB CACHING MODEL AND CORRESPONDING POLICIES

A. Modeling based on caching performance metrics

Perfect cache performance is the main motivation of replacement policies. When estimating performance, hit ratio, byte hit ratio and download latency are used commonly to measure performance of policies.

For the caching model above, we can estimate expected cost under policy A.

Let $\delta_i = \begin{cases} 1, & i\ is\ in\ cache; \\ 0, & else \end{cases}$, and the size of cache is M. The first user request was arrived at time 0.Thus the cumulative expected cost over the horizon [0, T] becomes

$$E_c(A) = E\left[\sum_{i=0}^{T} (1-\delta_i)\cdot c_i\right] \quad (2)$$

The expected average cost (over the infinite horizon) under policy A defined by

$$\overline{E_c}(A) = \lim_{T\to\infty} \frac{1}{T+1} E\left[\sum_{i=0}^{T} (1-\delta_i)\cdot c_i\right] \quad (3)$$

A number of situations can be handled by adequately

specializing the cost-per-step c_i, i.e., if c_i=1, good hit rate can be achieved; on the other hand, if c_i is taken to be the byte size s_i, then byte hit ratio is denoted. Therefore we have performance metrics as follows.

(1) Hit ratio (HR): Let c_i =1， i=1,…,N. Hit ratio under policy A is defined by

$$\mathrm{HR}(\mathrm{A})=\lim_{T\to\infty}\frac{1}{T+1}\mathrm{E}\left[\sum_{i=0}^{T}\delta_i\right] \tag{4}$$

(2) Byte hit ratio (BHR): Let c_i denote size of object i, $c_i = s_i$ .Byte hit ratio under policy A can be defined by

$$\mathrm{BHR}(\mathrm{A})=\frac{\lim_{T\to\infty}\frac{1}{T+1}\mathrm{E}\left[\sum_{i=0}^{T}\delta_i\cdot s_i\right]}{\lim_{T\to\infty}\frac{1}{T+1}\mathrm{E}\left[\sum_{i=0}^{T}s_i\right]} \tag{5}$$

(3) Download latency (LR): Another performance metric of great interest is the user-perceived download latency. Let c_i denote the delay fetching document i from Web server, is l_i, then download latency under policy A is defined by

$$\mathrm{LR}(\mathrm{A})=\lim_{T\to\infty}\frac{1}{T+1}\mathrm{E}\left[\sum_{i=0}^{T}(1-\delta_i)\cdot l_i\right] \tag{6}$$

In actual Web environment, there is conflict among different performance metrics. For example, hit ratio emphasizes particularly on reducing respond time, whereas byte hit ratio pays more attention to bandwidth spending. Since objects in cache have variable sizes, keeping more documents with small size can improve hit ratio, however, preferable byte hit ratio is not always obtained. On the other hand, saving large size documents can improve byte hit ratio. Network users always prefer reducing bandwidth, thus maximizing byte hit ratio is more important for them. In a word, we hope to find optimal policies under these factors to be pursued in systematic manner.

B. Modeling based on prefetching performance metrics

This subsection surveys the web performance indexes appeared in the open literature focusing on prefetch aspects. To the better understanding of the meaning of those indexes, we classify them into three main categories, according to the system feature they evaluate: 1) prediction related indexes. 2) resource usage indexes. 3) end-to-end perceived latency indexes. Surveys show that,indexes between caching and prefetching have some relations like:

$$L_{pre}=L_{cache}-\sum_{i=1}^{r}\frac{p_iT}{a_iu_i+1} \tag{7}$$

$$H_{pre}=H_{cache}+\sum_{i=1}^{r}\frac{p_i}{a_iu_i+1} \tag{8}$$

$$B_{pre}=B_{cache}+\sum_{i=1}^{r}\frac{s_i}{a_iu_i+1} \tag{9}$$

Where L_{cache}, H_{cache}, B_{cache} and L_{pre}, H_{pre}, B_{pre} denote the latency, hit ratio and byte hit ratio of caching and prefetching, respectively. These formulae show that caching and prefetching have the common optimizing aim that the one with lower cost, smaller size and more accessing frequency.

C. Optimal model

Analysis in chapter 2 shows that, conventional caching model fails to capture temporary locality and statistical information contained in the stream of requests. To describe user request sequences more exactly and improve performance of Web cache and prefetching, Relative Popularity and Byte Cost are presented to optimize Web caching and prefetching model.

Definition1: Relative Popularity (RP_i): rate of popularity of each document and the highest popularity.

$$\mathrm{RP}_i=\frac{\mathrm{P_i}}{\mathrm{D}} \tag{10}$$

D is a parameter means the most popular Web objects. We can obtain its value through normalized computing.

Definition2: Byte Cost (U_i): cost of unit byte.

$$\mathrm{U}_i=\frac{\mathrm{c_i}}{\mathrm{s_i}} \tag{11}$$

When cache is full, objects with low cost (i.e., less popular, low latency, large size) are evicted out from cache.

Let optimal cost at time k be c_k. After R_k arrived, the state of cache becomes $S_{k+1}=S_k-V_k+R_k$. Thus optimal cost at time k+1 is

$$\begin{aligned}\mathrm{c}_{k+1}&=\mathrm{c}_k+\sum_{i\in V_k}\mathrm{RP}_i\cdot\mathrm{U}_i\\&=\mathrm{c}_k+\sum_{i\in V_k}\mathrm{RP}_i\cdot\mathrm{U}_i\end{aligned} \tag{12}$$

Since we can't control c_k, to optimize the cost, minimal $\sum_{i\in V_k}\mathrm{RP}_i\cdot\mathrm{U}_i$ should be selected. We can achieve a Web caching and prefetching optimal model called PR PPM based on RP_i and U_i as follows.

Input: R= {R_1, R_2, …, R_n}
Output: Caching objects
Method:
Step1. if R_t is in cache
Update U_i

Figure 1. Algorithm for replacement in PR PPM

Input: user sequence LF
Output: prediction model T
Method:
Step1. initialize, T=Null
Step2. for each R_t in request sessions, `create T which contains URL and other accessing record`
Step3. return T

Figure 2. Algorithm for prediction in PR PPM

D. Simulations

Request trace forms Zipf's law, and we select parameter a=0.75 in this experiment. Simulation is based on actual log: USASK-HTTP[6]. Representative PPM models like Standard PPM, LPS PPM and PB PPM are selected to compare with PR PPM.

Consider hit ratio, byte hit ratio precision and recall as indexes for performance evaluation given by

$$HR = H / R \tag{13}$$

$$BHR = H_B / R_B \tag{14}$$

$$\text{Pr}\,ecision = p^+ / p \tag{15}$$

$$\text{Re}\,call = p^- / R \tag{16}$$

We define below some basic concepts that have been used in above formulae.

H: amount of requests hit.

R: amount of user requests.

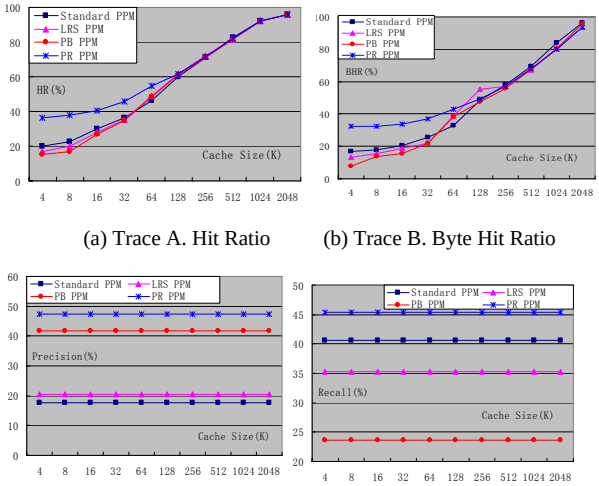
H_B: amount of hit requests bytes.

R_B: amount of requests bytes.

p: amount of objects predicted by the prediction engine.

p^+: amount of prefetched objects that are subsequently demanded by the user.

p-: amount of objects prefetched by the prefetching engine.



(a) Trace A. Hit Ratio (b) Trace B. Byte Hit Ratio

(c) Trace C. Precision (d) Trace D. Recall

Figure 3. Simulations on USASK-HTTP

Figure 3 shows the result of PR PPM and other PPM Models: Standard PPM, LPS PPM and PB PPM in different performance metrics. PR PPM optimizes latency in caching algorithm of this model which caters to the user's point of view. The figures show that, PR PPM has better hit ratio, byte hit ratio because it can reflect the transfer of users' interests. At the same time, PR PPM performances well in precision and recall as the close relation among different categories that high hit rate can make good precision, and also the cost decrease.

III. Conclusions

PR PPM is studied in this paper. By analyzing advantages and weaknesses of conventional caching model and prefetching model, this paper presents optimal Web caching and prefetching model named PR PPM based on the measurement of Relative Popularity and Byte Cost. PR PPM order Web objects by size, value and download latency in systematic manner. And also PR PPM has better hit ratio, byte hit ratio because it can reflect the transfer of users' interests. Therefore, good hit ratio and byte hit ratio are achieved and total precision and recall is optimal at the same time.

References

[1] L. Xiao, X. Chen, X. Zhang, etc. On scalable and locality-aware Web document sharing. J. of Parallel and Distributed Computing, 2003, 63(10):945-962.

[2] D. Starobinski, D. Tse. Probabilistic methods for Web caching. Performance Evaluation, 2001, 46 (2-3): 125−137.

[3] K. Psounis, A. Zhu, B. Prabhakar, etc. Modeling correlations in web traces and implications for designing replacement policies. Computer Networks, 2004, 45(4): 379− 398.

[4] O. Bahat and A.M. Makowski. Optimal replacement policies for non-uniform cache objects with optional eviction. San Francisco: In Proceedings of INFOCOM 2003, April 2003:427-437.

[5] Lei Shi, Zhimin Gu, Lin We, etc. An Applicative Study of Zipf's Law on Web Cache. International Journal of Information Technology, 2006, 12(4):49-58.

[6] http://www.usask.ca/

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 254-257

Palmprint Recognition Based on Gabor Transforms and Invariant Moments*

Rina Su, Yongping Zhang, Jianbo Fan, and Shaojing Fan
School of Electronic and Information Engineering
Ningbo University of Technology，Ningbo, China
srn2009@126.com,zhangyp1963@yahoo.com, jbfan@nbut.cn,fsj@nbut.cn

***Abstract*—Proposed a method of combining two-dimensional Gabor transforms and invariant moments to extract palmprint feature, and using multilayer towards feedback neural network for training palmprint images to recognize. This method first pretreated the collected palmprint images and got the region of interest (ROI), then constructed a set of Gabor filters to get ROI eigenvectors, combined with the palm images' invariant moment features together as the input of the neural network to train and recognize. Experiments show that the method is effective.**

***Index Terms*—palmprint recognition, Gabor transforms, invariant moments**

I. INTRODUCTION

Palmprint identification is a biometric identification technology researching in recent years. It uses image-processing and pattern-recognition methods by analyzing the person's palm for identification. Compared with the fingerprints, iris, face and other biometric recognition technology, palmprint identification has some advantages. For example its main characteristics are stable and obvious, and the noise interference is less [1], so palmprint biometric identification has become a research hotspot. It can be widely used in access control, forensic, medical and social security, information systems security and other fields. Currently, the palmprint recognition methods include space structural-based features, frequency domain-based features, and statistical-based features for identification.

Space structural-based features mainly analysis the distribution characteristics on the palm including the main line, wrinkles, minutiae and other characteristics; frequency domain-based features mainly transform the original images from space domain to frequency domain, and extract its features for description the palmprint and analysis. This method mainly includes wavelet transform, Fourier transform, Gabor transform, etc.; statistical-based features for recognition uses the statistical methods for re-definition and measurement the original images, mainly including the invariant moment characteristics.

Gabor transform is one of the frequency domain-based analysis methods, and it is a powerful tool for texture analysis. Invariant moment describes the shape of the image features, and it is also widely used in the field of image feature extraction. Because the individual differences of the palms are mainly in two aspects of different textures and different shapes, so in this paper, we propose to combine the Gabor transform and invariant moment eigenvector for palmprint identification. Experiments prove that this method has a high recognition rate by comparing with other methods.

II. PALMPRINT RECOGNITION SYSTEM

Palmprint recognition usually includes the processes of palm image acquisition, preprocessing, feature extraction, classification and identification. Palmprint image acquisition primarily completes tasks of acquisition and preservation the original images, and then through the pre-processing stage to carry out the operation of removing noises, enhancement, segmentation, positioning and normalized, and form the standard palm database, then calculate and extract the images' feature information in the database. Some of the palmprint data form palm template database, and the others form the samples to be identified. Finally, according to the extracted feature information and the matching algorithm identify the palmprint.

A. image acquisition and pretreament

We use flatbed scanner as gathering tool for palmprint images. Scanner's model is Samsung SCX-4725. Choose 30 persons to scan the palmprint of their right hands; every person is collected five palmprint images, the 150 palmprint images with 710 × 279mm and 256-level BMP form database. Because the collected image data is large and the same person' different palm images may have different degrees of translation and rotation, it is not conducive to extract feature and recognize. Therefore, it needs to go through the pretreatment process to get the ROI regions. In this paper, we reference the approach in paper[2], that is, firstly interception the original image into 398 × 279 size with 256bmp format images which need to contain palm main part, and then convert the intercept part into binary images; And then automatic detect two anchor points A (the bottom point of the vertical axis on the index finger and middle finger gap) and B (the bottom point of the vertical axis on the little finger and ring finger gap); Establish rectangular coordinate system by the two anchor points(A and B) connection line and the midpoint on the connection line; Rotate the original palmprint images into the new

* This work is supported by Zhejiang Province Natural Science Foundation #Y1080123 and overseas student science and technology activities preferred funded projects(Department of Human Resources and Social Security,2009).

AP-PROC-CS-09CN005

rectangular coordinate system; Cut palm size of 128 × 128 sub-images to form palm ROI of the standard database. The sample of ROI database is shown as Fig.1.

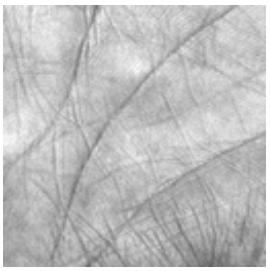


Figure 1. palmprint ROI

B. Feature Extraction

(1)Gabor transform

In order to study the local frequency characteristics signal, Dennis Gabor proposed Famous "Window" Fourier Transform (Also known as short-time Fourier transform, STFT) in the paper "Theory of communication" in 1946，later called Gabor transforms. The Gabor transform function *f(t)* defined as:

$$G_f(\omega,\tau)=\int_{-\infty}^{\infty} f(t)g(t-\tau)e^{-j\omega t}dt \quad . \qquad (1)$$

$g(t)$ is an appropriate window function, for example, Gaussian function. We can also consider that the base function of Gabor transform is equivalent to the modulation function of window function, which is Gabor function. Studies have shown that Gabor function is the only function to achieve the lower bound of uncertainty relation. Therefore, it can achieve the best localization in time - frequency domain at the same time.

In 1985, Daugman made Gabor function expanded into two-dimensional form in paper [4], and on this basis constructed a 2D Gabor filter. Two-dimensional Gabor filter can be seen as the positive spin wave plane modulated by a two-dimensional Gaussian function with specific frequency and direction. The general form of 2D Gabor filter basis functions is:

$$\psi_{m,n}(x,y)=\alpha_0^{-m}\exp\left\{-\frac{\alpha_0^{-2m}}{8}\left[4\left(x\cos\frac{n\pi}{K}+y\sin\frac{n\pi}{K}\right)^2+\left(-x\sin\frac{n\pi}{K}+y\sin\frac{n\pi}{K}\right)^2\right]\right\}$$

$$\times\exp\left[i\omega\alpha_0^{-m}\left(x\cos\frac{n\pi}{K}+y\sin\frac{n\pi}{K}\right)\right] . \qquad (2)$$

ω is oscillation frequency, α_0^{-m} is the scale factor, and K is on behalf of the overall direction number of filters. By selecting different values of m and n, we can construct a set of Gabor filters in different scales and directions. We can realize Gabor transform by convolution Gabor filters with the original image, the result eigenvectors can be used to detect and describe the texture features of the image.

Equation (2) can be divided into real and imaginary parts and they can also be separately applied into the image feature extraction. For the front ROI images after pretreatment, named I(x,y), we use $w_{m,n}(x, y)$ and $w^*_{m,n}(x, y)$ represent the Gabor filtering transform and the conjugate transform, then we can get the sub band in the m-scale and the n-direction:

$$\begin{aligned}\overline{w}_{m,n}(x, y)&=\frac{1}{2}\left[w_{m,n}(x, y)+w^*_{m,n}(x, y)\right]\\&=\iint I(s,t)\overline{\psi}(x-s, y-t)dsdt.\end{aligned} \qquad (3)$$

$\overline{\psi}(x-s, y-t)$ represents the real part of Gabor filter.

In order to maintain rotation invariance of the palmprint images, we use average value in different directions at the same scale to describe the texture, so we define:

$$Gf_m(x, y)=\frac{1}{K}\sum_{n=0}^{K-1}\overline{W}_{m,n}(x, y) \quad . \qquad (4)$$

We need to calculate the mean of $Gf_m(x, y)$ named μ_m and its standard deviation named σ_m to express texture features, and then form a texture eigenvector:

$$\overline{v}=(\mu_0, \sigma_0, \mu_1, \sigma_1,..., \mu_{s-1}, \sigma_{s-1}).$$

The level of recognition rate is limited by the expression of characteristics, so it is important to select reasonable parameters of scale and direction. There have not yet uniform effective standards for how to select the parameters, but according to the constant experimental test summary. In this paper, we select 8 directions with $\pi\backslash 8$ interval, that is n=1,2,3,4,5,6,7(K=8), with the samples in different directions, we select 5 scales, that is m=0,1,2,3,4;We construct a group of 5*8-dimensional filters, and then calculate $\overline{w}_{m,n}(x, y)$, according to equation (4),get the mean and variance of $Gf_m(x, y)$, that is :

$$\overline{v}=(\mu_0, \sigma_0, \mu_1, \sigma_1, \mu_2, \sigma_2, \mu_3, \sigma_3, \mu_4, \sigma_4).$$

(2) Moment invariants

The palmprint image may generate deformation because of the subjective and objective factors in the acquisition process. Although the pre-processing stage can adjust the images, but it can not completely eliminate deformation. M.K.Hu first proposed the definition of continuous function moments in 1962[6] for image description, and given seven moment invariant expressions with features of translation, scaling and rotation invariance.

Therefore, in this paper, we consider to extract the feature with the graphics invariant feature, avoid the error caused by deformation at rough matching stage. At the same time, in order to avoid the high dimension of eigenvectors, we take the first two moments in Hu moments as the moment invariant of palm ROI region.

R is expressed as the ROI region, and the moment of R is as follows:

$$M_{ij}(R)=\sum_{(x,\ y)\in R} x^i y^j \ . \tag{5}$$

M_{00} is the number of points in the region R, and it represents the area of R. The center of R is expressed as:

$$(\bar{x},\bar{y})=(M_{10}/M_{00},M_{01}/M_{00}) \ . \tag{6}$$

Then we can calculate the central moments of R region:

$$m_{ij}(R)=\sum_{(x,\ y)\in R} (x-\bar{x})^i (y-\bar{y})^j \ . \tag{7}$$

Therefore, we can calculate Normalized Hu moments:

$$M_1=(m_{20}+m_{02})/m^2_{00} \tag{8}$$

$$M_2=\left[(m_{20}-m_{02})^2+4m^2_{11}\right]/m^2_{00} \ . \tag{9}$$

C. Classification and recognition

Neural networks have the function of self-organization and adaptive learning. As long as the model can be identified with certain differences, the network can identify different models through adaptive clustering, so it is widely used in the field of pattern recognition. There are various models of neural networks and they describe and simulate nervous system in different angles and levels.

In this paper, we use back-propagation learning algorithm (BP algorithm) of neural network. It is composed of two processes by the signal being transmitted forward and the error back-propagation. The input samples are imported into the input layer in forward propagation, through each hidden layer disposed layer by layer, the signal is transmitted to the output layer. If the actual output does not match the expected output, then it is transferred to error back-propagation phase. Back-propagation, the output errors are transmitted to input layer through hidden layer and errors are apportioned to all the units, this error signal can be used to amended value. The process of signal transmitted and error back-propagation is carried out roundly. Constantly the weight is adjusted; this is the network learning training process.

For each palm images in the training sample space, we form a 12-dimension eigenvector $(M_1,\ M_2,\ \mu_0,\ \sigma_0,\ \ldots,\mu_4,\ \sigma_4)$ with two central moments calculated before and 10 Gabor features at 5 scale, and then built supervised classification and recognition systems based on BP neural network, design the input layer nodes is 12. So each component of the 12-dimensional eigenvector corresponds to an input node, and then design a hidden layer as a classifier with the input palmprint eigenvectors, based on Klmogorov theory (that neural network input layer nodes are N , then the hidden layer nodes set to 2N +1), so the number of nodes in hidden layer is defined as 25. The output layer nodes are equal to the number of categories to be matched palm. Mathematical neural network model is expressed as:

$$out_i=\varphi\left(\sum_j w_{ij}out_j+\theta_i\right) . \tag{10}$$

Here, out_i represent the current output of the i-nodes, out_j on behalf of the former layer of the output of the first j- nodes, φ on behalf of a non-linear function, such as our selection $\varphi(t)=1/(1+e^{-t})$, and then we use BP algorithm for training.

III. EXPERIMENTAL RESULTS

We collected 30 persons' right hands palm images, every person is 5 and total is 150. Then we get every person's each palm image as a template (total 30). The remaining 120 are as the training samples. In our experiment, we choose the experiment scale parameter as 0, 1,2,3,4. Then we use the method described above to calculate the eigenvector $(M_1,\ M_1,\ \mu_0,\ \sigma_0,\ldots,\ \mu_4,\ \sigma_4)$ as the input of the neural network to test. The classification results shown in Table Ⅰ, in the experiment, the correct recognition rate can achieve 96.7%. It is higher than the method of only use identification-based on Gabor transforms or digital invariant moments.

TABLE I. COMPARISON

Categories	Sample number	Error number	Rejection number	Recognition rate
Identification method based on Gabor Transform	120	3	5	93.3%
Invariant Moment Recognition	120	5	7	90.0%
This method	120	2	2	96.7%

Ⅳ. CONCLUSION

This article give a method of combining Gabor transforms and the invariant moments to extract palmprint images' eigenvector, and then build the BP neural network to train and identify the palm, in the experiments we select the appropriate threshold for self-collected palmprint database to test. Compared with the Gabor transform-based method and the moment-based method, this method can obtain higher recognition rate.

REFERENCES

[1]Ping Xie, Zhifeng Zhou. Palmprint feature extraction based on wavelet transform and entropy [J]. System Application of Computer, 2008, 2:105-107.

[2]Meng Wang, Qiuqi Ruan.Capture and Preprocess of Palmprint Image[J].Application Research of Computers，2007，6：161-164.

[3]D.Gabor.Theory of communication.Journal of Institute for Electrical Engineering.93:429-457，1946.

[4]J.Daugrnan.Uncertainty relation for resolution in space，spatial frequency and orientation optimized by two-dimensional visual cortical filters.Journal of the Optical Society of America A.2:1160-1169，1985.

[5]Y.Zhang,D.W.Fountain,R.M.Hodgson,J.R.Flenley,S.Gunetileke.Towards automation of palynology 3:pollen pattern recognition using Gabor transforms and digital moments.Journal of Quaternary Science.19(8)763-768，2004.

[6]Hu M K.Visual Pattern Recognition by Moment Invariants[J].IRETrans.Information theory,1962,IT(8):179-187

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 258-261

Convexity Conditions of Planar Parametric Curves and its Properties

Kui Fang [1], Xinghui Zhu [2], Wu Luo [2] , Juan Wang [2] , and Yujuan Wang [2]
[1] School of Information Science & Technology, Hunan Agricultural University , Chang Sha , China, 410128
Email: fk@hunau.net
[2] School of Information Science & Technology, Hunan Agricultural University , Chang Sha , China, 410128
Email: wang_126juan@126.com

***Abstract*—Based on the original geometrical definition of the planar parametric convex curve, the local convexity and the global convexity on parametric curve are discussed, a few properties are obtained, and a necessary and sufficient condition is presented for the local convexity of a general planar parametric curve.**

***Index Terms*—planar parametric curves, convexity, Relative curvature**

I. INTRODUCTION

The convexity of parametric curve and surface is an intuitive geometry concept in mathematics. We also have known that the convexity of single value functions can be clearly defined by second derivatives and partial derivatives. In the literatures on Computer Aided Geometric Design (CAGD), this concept of convexity has been described in various ways, but it used without an explicit definition.

The convexity was occurred and studied in Differential Geometry, Functional Analysis and Integral Geometry. In addition, its result has been used in CAGD. In Functional Analysis only the convexity of single-value function is considered, and in differential and integral Geometry , mainly the convexity of planer, simple, closed curves is investigated. In CAGD the researches to convexity of parametric curves and surfaces have more subjects and more useful in practice. For example, parametric curve can be classed as opening curve and closed curve according to its type, and with respect to its property, it can be classed as single-value and non-single-value curve. It is not evident that the above concepts of convexity can be used easily in CAGD.

Many important results to research of convexity have been given in CAGD. Such as the convexity of Bézier curve by Su Bu Qing Ref. [4] and the convexity of B-spline curves by Liang You Dong Ref. [5] and de Boor Ref. [6]. However, these results are connected with special some methods, therefore, it is not applicable to general curves.

The research described in this paper was the convexity of parametric curve. The parametric curve is convex if there are no inflection points in it. It's sufficient in practice that there is only convexity concept. For example, the curve is convex if it has self-intersection points (twist point), but it's not suitable to apply in practice. So it is necessary to class convexity as local convexity and global convexity. In Wang Ref. [7] and C.Liu Ref. [2] , the relative curvature of curve is non-negative (non-positive) is defined directly as the locally convex. In this paper, based on the original geometrical definition of the convex curve, the local and global convexity of curve are discussed , a few properties are obtained, the relative curvature of curve is non-negative (non-positive) , which is proved.

Corresponding author: Tel.: +0086-0731-84635363
E-mail address:Fk@hunau.net

Ⅱ. DEFINITION OF CONVEX CURVE

A. Definition 2.1

A planar oriented curve is an ordered set in R2 , given by

$$\Gamma : \mathbf{r}(t) = (x(t), y(t)), a \le t \le b$$

Where the parameter t from a to b. If the curve is counterclockwise, using a transformation $t = b + (t-a)/(b-a)$, then the curve becomes clockwise, and we assumed that the curves in this paper is always clockwise.

B. Definition 2.2

Given parametric curve Γ **:** $\boldsymbol{r}(t)$, $a \le t \le b$,if

$$\boldsymbol{r}'(t) \ne 0, a \le t \le b$$

Then, we call Γ as regular parametric curve. In this paper, the regular parametric curves are considered only.

Given the direction of tangent vector of Γ is consistent with the parameter t , so the direction of tangent vector is defined as the direction of the tangent line, and the plane on which the tangent line lies is divided into two half planes by the tangent, and along the direction of the tangent, the half-plane on the right side of tangent line is called the right half-plane, and the other half-plane is called the left half-plane.

The global convexity of a planar curve is defined as follows:

C Definition 2.3

Given a regular planar curve $\Gamma : r = r(t)$, $a \le t \le b$, $\forall t \subset [a,b]$, if the entire curve Γ lies in a side of the tangent line of Γ at point $P = r(t)$, then P is called as a globally convex point.

D Definition 2.4

Given a regular planar curve $\Gamma:\mathbf{r}=\mathbf{r}(t)$, $a\le t\le b$,if every points on Γ are all globally convex point , then Γ is called as a globally convex curve.

Suppose that the normal vector of the plane in which Γ is lying is an unit vectork, the unit tangent vector of Γ is $\boldsymbol{\alpha}(t)$ at $P=\mathbf{r}(t)$, and $\overline{P}=\mathbf{r}(\bar{t})$ is an arbitrary point on Γ, Let

$$e(t,\overline{P})=[\boldsymbol{\alpha}(t)\times\overrightarrow{P\overline{P}}]\cdot\mathbf{k}\qquad a\le t\le b \tag{1}$$

It is clear that $\overline{P}$ is on the closed left (right) half-plane of the tangent line of Γ, if and only if

$$e(t,\overline{P})\ge 0(\le 0)\qquad a\le t\le b \tag{2}$$

Thus we have the following theorem:

E Theorem 2.1

A regular planar curve $\Gamma:\mathbf{r}=\mathbf{r}(t)$, $a\le t\le b$, then Γ is a globally convex curve, if and only if that Γ lies on the closed left (right) half-plane of the tangent line of an arbitrary point on Γ.

Proof Suppose that Γ is in the closed right half-plane of tangent at the initial point $P_s=\mathbf{r}(a)$, and satisfying

$$e(a,\overline{P})=[\boldsymbol{\alpha}(a)\times\overrightarrow{P_s\overline{P}}]\cdot\mathbf{k}\ge 0$$

Because Γ is a regular curve, for the fixed point $\overline{P}$, $e(t,\overline{P})$ is a continuous function. If the sign of $e(t,\overline{P})$ is changed at $t=t^*>a$, that is, when $t\le t^*$, Γ is in the closed right half-plane of the tangent at the point $P=\mathbf{r}(t)$, then

$$e(t,\overline{P})\ge 0\,,e(t^*,\overline{P})=0 \tag{3}$$

And there exists a ε, when $t^*<t<t^*+\varepsilon$, have $e(t,\overline{P})<0$

According to (1), we know that $e(t^*,\overline{P})=0$ if and only if $\overline{P}$ lies on the tangent at $P^*=\mathbf{r}(t^*)$. By the (3), all points on Γ are in the closed right half-plane of the tangent at P^*, thus as long as Γ isn't a straight line, then there exists a point P_1 on Γ, shch that P_1 isn't in the tangent line at $P^*=\mathbf{r}(t^*)$, and

$$e(t^*,P_1)=[\boldsymbol{\alpha}(t^*)\times\overrightarrow{P^*P_1}]\cdot\mathbf{k}>0$$

Obviously, $e(t,P_1)$ is also a continuous function, thus there exists a neighborhood of t^*, i.e. $t^*<t<t^*+\delta$, in the neighborhood ,with

$$e(t,P_1)=[\boldsymbol{\alpha}(t)\times\overrightarrow{PP_1}]\cdot\mathbf{k}>0$$

Therefore, there exists a neighborhood of t^*, $t^*<t<t^*+\sigma$, with $\sigma=\min(\varepsilon,\delta)$, satisfying $e(t,\overline{P})<0$, $e(t,P_1)>0$

According to (2), the above inequality explains that when $t^*<t<t^*+\sigma$, $\overline{P}$ and P_1 are in the left and right half-planes of tangents separately, which contradicts with that Γ is a globally convex curve , so the sign of $e(t,\overline{P})$ is unchanged,

So $\overline{P}=r(\bar{t})$ is always in the closed right half-plane of the tangent of Γ .because $\overline{P}=r(\bar{t})$ is arbitrary point, Γ is in the closed right half-plane of the tangent at any point P.

When Γ is in the left half-plane of the tangent at the initial point $P_s=\mathbf{r}(a)$, the method of proof is as same as the method above. Necessity is proved. The theorem is proved.

The converse, the sufficiency can be proved directly by the definition of globally convex curve.

F Defitition2.5

Given a regular planar curve $\Gamma:\mathbf{r}=\mathbf{r}(t)$, $a\le t\le b$, $\forall t\subset[a,b]$, if there exists a neighborhood $\delta(t,\varepsilon)$ of $P=\mathbf{r}(t)$, in the neighborhood, the corresponding curve segment on Γ completely lies in a side of the tangent line at point P, then P is called as a globally convex point.

G Definition 2.6

Given a regular curve $\Gamma:\mathbf{r}=\mathbf{r}(t)$, $a\le t\le b$, $\forall t\subset[a,b]$, if there exists a neighborhood $\delta(t,\varepsilon)$ of $P=r(t)$, in the neighborhood, the corresponding curve segment on Γ completely lies in the right (left) half-plane of the tangent line at the point P, then Γ is called as a locally convex curve.

Each point in locally convex curves is locally convex point. However, although each point in Γ is locally convex point, the curve isn't locally convex curve.

III.THE PROPERTIES OF CONVEX CURVES

To discuss conveniently, we use arc-lenth parametric curves. Assume $\mathbf{k}$ is an unit vector which points to the top of the plane that Γ lies in , noting $\mathbf{n}=\mathbf{k}\times\boldsymbol{\alpha}$, then $\mathbf{n}$ always points to the left half-plane of tangents. When $\boldsymbol{\beta}$ is replaced by $\mathbf{n}$, the Frenet formula of Γ as follows

$$\begin{cases} d\boldsymbol{\alpha}\ /\ ds\ =\ \kappa_r(s)\,\mathbf{n} \\ d\mathbf{n}\ /\ ds\ =\ -\kappa_r(s)\,\boldsymbol{\alpha} \end{cases}$$

Where $\boldsymbol{\alpha},\kappa_r(s)$ are the tangent vector and the relative curvature of Γ respectively.

We do the Taylar expansion at P, with

$$\mathbf{r}(s+\Delta s)-\mathbf{r}(s)=\dot{\mathbf{r}}\Delta s+\frac{1}{2!}(\ddot{\mathbf{r}}+\boldsymbol{\varepsilon}_1)(\Delta s)^2 \tag{4}$$

$$\mathbf{r}(s+\Delta s)-\mathbf{r}(s)=\dot{\mathbf{r}}\Delta s+\frac{1}{2!}\ddot{\mathbf{r}}(\Delta s)^2+\frac{1}{3!}(\dddot{\mathbf{r}}+\boldsymbol{\varepsilon}_2)(\Delta s)^3 \tag{5}$$

And $\boldsymbol{\varepsilon}_i = \varepsilon_i'\boldsymbol{\alpha} + \varepsilon_i''\boldsymbol{\beta}(i=1,2)$ $\lim_{\Delta s \to o} \boldsymbol{\varepsilon}_i = \mathbf{0}$.

According to the Frenet-formular, have $\dot{\mathbf{r}} = \boldsymbol{\alpha}$, $\ddot{\mathbf{r}} = \kappa_r \mathbf{n}$, $\dddot{\mathbf{r}} = \dot{\kappa}_r \mathbf{n} + \kappa_r \dot{\mathbf{n}} = \dot{\kappa}_r \mathbf{n} - \kappa_r^2 \boldsymbol{\alpha}$

When $\dot{\mathbf{r}}$ 、 $\ddot{\mathbf{r}}$ 、 $\dddot{\mathbf{r}}$ are substituted into the equation (4) and (5), we have

$$\mathbf{r}(s+\Delta s) - \mathbf{r}(s) = [\Delta s + \frac{1}{2}\boldsymbol{\varepsilon}_1'(\Delta s)^2]\boldsymbol{\alpha} + \frac{1}{2}(\kappa_r + \boldsymbol{\varepsilon}_1'')(\Delta s)^2 \mathbf{n} \qquad (6)$$

$$\mathbf{r}(s+\Delta s) - \mathbf{r}(s) = [\Delta s + \frac{1}{6}(-\kappa_r^2 + \boldsymbol{\varepsilon}_2')(\Delta s)^3]\boldsymbol{\alpha} + [\frac{1}{2}\kappa_r(\Delta s)^2 + \frac{1}{6}(\dot{\kappa}_r + \boldsymbol{\varepsilon}_2'')(\Delta s)^3]\mathbf{n} \qquad (7)$$

Thus

$$[\mathbf{r}(s+\Delta s) - \mathbf{r}(s)] \cdot \mathbf{n} = \frac{1}{2}(\kappa_r + \boldsymbol{\varepsilon}_1'')(\Delta s)^2 \qquad (8)$$

$$[\mathbf{r}(s+\Delta s) - \mathbf{r}(s)] \cdot \mathbf{n} = [\frac{1}{2}\kappa_r(\Delta s)^2 + \frac{1}{6}(\dot{\kappa}_r + \boldsymbol{\varepsilon}_2'')(\Delta s)^3] \qquad (9)$$

A Theorem 3.1

Given a planar regular curve $\Gamma : \mathbf{r} = \mathbf{r}(t)$, $a \le t \le b$, if the P isn't a stationary point in Γ ,then there exists a neighborhood of P ,when $\kappa_r > 0$, in the neighborhood, the corresponding curve segment is in the closed left half-plane of the tangent at P , when $\kappa_r < 0$, in the closed right half-plane.

Proof. Suppose that P isn't the stationary point on Γ , that is $\kappa_r \neq 0$ at P , Γ is represented as arc-lenth parametric curve. When $\kappa_r > 0$, in (7) , if Δs is small enough , existing $\kappa_r + \boldsymbol{\varepsilon}_1'' > 0$, then

$$[\mathbf{r}(s+\Delta s) - \mathbf{r}(s_0)] \cdot \mathbf{n} > 0$$

By the same reason , when $\kappa_r < 0$, as long as Δs is small enough , existing $\kappa_r + \boldsymbol{\varepsilon}_1'' < 0$, then

$$[\mathbf{r}(s+\Delta s) - \mathbf{r}(s_0)] \cdot \mathbf{n} < 0$$

Hence a neighborhood of P always exists , when $\kappa_r > 0$, and the corresponding curve segment is in the closed left half-plane of the tangent at P ; when $\kappa_r < 0$, in the closed right half-plane . The theorem is proved.

B Theorem 3.2

Given planar regular curve $\Gamma : \mathbf{r} = \mathbf{r}(t)$, $a \le t \le b$, suppose that P is a stationary point in Γ , if $\kappa_r' \neq 0$, then P is n′ t a locally convex point.

Proof. If P is a stationary point in Γ , and $\kappa_r' \neq 0$ at P , then $\dot{\kappa}_r = \kappa_r' / |\mathbf{r}'| \neq 0$. To discuss conveniently, we assume $\kappa_r > 0$, and if $|\Delta s|$ is small enough, having

$$\dot{\kappa}_r + \boldsymbol{\varepsilon}_2'' > 0$$

In the equality (3.8) , when $\Delta s > 0$, then

$$[\mathbf{r}(s+\Delta s) - \mathbf{r}(s)] \cdot \mathbf{n} > 0$$

However, when $\Delta s < 0$, having

$$[\mathbf{r}(s+\Delta s) - \mathbf{r}(s)] \cdot \mathbf{n} < 0$$

This explains that P is not a locally convex point. The proof of theorem is completed.

C Lemma 3.1

A planar regular curve $\Gamma : \boldsymbol{r} = \boldsymbol{r}(t), a \le t \le b$, suppose that θ is the oriented included angle formed by the tangent vector $\boldsymbol{\alpha}$ at P in Γ and the direction vector of $x-$axis, then

$$d\theta(t)/dt = k_r(s)\left|d\mathbf{r}/dt\right|$$

Especially when Γ is arc-lenth parametric curve , existing

$$d\theta(s)/ds = k_r(s)$$

D Lemma 3.2

A planar regular curve $\Gamma : \boldsymbol{r} = \boldsymbol{r}(t), a \le t \le b$, then there exists a point in Γ , its tangent vector is parallel to $\boldsymbol{r}(b) - \boldsymbol{r}(a)$.

Proof. If $\boldsymbol{r}(a)$ is defined as the origin point , and the direction of $\boldsymbol{r}(b) - \mathbf{r}(a)$ is the direction of the x - axis, noting

$$\boldsymbol{r} = \boldsymbol{r}(x) = \{x, f(x)\}$$

When Γ is a regular curve, $f(x)$ is continuously differentiable. Suppose that x^* is the local extremum value of $f(x)$, then $f'(x^*) = 0$, and the tangent vector of the curve at x^* is

$$\boldsymbol{r}'(x^*) = \{1, 0\}$$

Thus $\boldsymbol{r}'(x^*)$ parallels to $\boldsymbol{r}(b) - \boldsymbol{r}(a)$. The lemma is proved.

E Theorem 3.3

A planar regular curve Γ is a locally convex curve , if and only if the relative curvature κ_r is non-positive (non-negative).

Proof. Suppose that Γ is arc-lenth parametric curve, if Γ is a locally convex curve, for any point P in Γ ,a neighborhood $\delta(s,\varepsilon)$ of s exists ,when $s \subset \delta(s,\varepsilon)$, the corresponding curve segment of Γ completely lies in a side of the closed right half-plane of the tangent at P , given by (8)

$$(\kappa_r + \boldsymbol{\varepsilon}_1'')(\Delta s)^2 \ge 0 (\le 0)$$

Get limit on left sides of the equality above, obtaining

$$\kappa_r \ge 0 (\le 0)$$

So the sign of the relative curvature κ_r is unchanged, the necessity is proved.

Next, suppose $\kappa_r(s) \le 0$, for any point $\overline{P} = \mathbf{r}(\overline{s})$ of Γ, if $\kappa_r < 0$, then because of theorem 3.1, a neighborhood exists of $\overline{P}$,and the corresponding curve segment lies in the right half-plane of tangent at $\overline{P}$; If $\kappa_r = 0$, if a neighborhood above at $\overline{P}$, doesn't exist, then a point $P^* = \mathbf{r}(s^*)$ in Γ is surely lying in the left half-plane of tangent at $\overline{P}$.

If $s^* > \overline{s}$, according to the lemma 3.2, exist a point P_1 in the curve segment with the point $\overline{P}$ as the starting point and P^* as the a terminal point, the tangent vector $\boldsymbol{\alpha}_1$ at P_1 parallels to $\overrightarrow{\overline{P}P^*}$. Suppose that $\overline{\theta}$ is the included angle formed by the tangent vectors $\boldsymbol{\alpha}_1$ and $\overline{\boldsymbol{\alpha}}$ at $\overline{P}$, then $0 < \overline{\theta} < \pi$. Suppose that $\boldsymbol{r}(a)$ is defined as the origin point, and the direction of $\boldsymbol{\alpha}_1$ as the direction of the x - axis, if the included angle is θ formed by the tangent vector $\boldsymbol{\alpha}$ at P of Γ and the x - axis, according to the lemma 3.1, $d\theta / ds = \kappa_r \le 0$, that is, θ is monotone decreasing, which is in contradiction with the result, $0 < \overline{\theta} < \pi$.

In a similar way, the result can obtain.

To sum up, for any point $\overline{P}$ in Γ, there exists always a neighborhood of $\overline{P}$, such that the corresponding curve segment lies in the right half-plane of the tangent at $\overline{P}$, thus Γ is a locally convex curve.

When $\kappa_r(s) \ge 0$, the method on proof is completely similar. The theorem is proved.

Ⅳ .Conclusion

According to the original geometrical definition of planar parametric convex curve, the local convexity and the global convexity on parametric curve are discussed, and a few properties are presented. The necessary and sufficient conditions about the locally convexity of parametric curves are obtained. In addition, the relationship between inflection points and locally convexity for parametric curves will be investigated in next paper.

References

[1] K Wilhelm. (1978) , A Course in Differential Geometry，Spring-Verlag (in New york) .

[2] C.Liu and C.R.Trass, On convexity of planar Curves and its Application in CAGD,CAGD,Vol.14,No.6,653-669,1997.

[3] T.N.T. Goldman, Shape Preserving Representations, in Mathematical methods in CAGD(eds. T.Lyche and L.L.Schumaker), Academic Press, Boston, 1989,333-357

[4] S.Buqin, LiuDingyuan (1981), Computer Geometry , ShangHai Science And Technology Publishing House.

[5] L.Youdong. (1982) , Theorem of B－Spline geometric curve and surfaces and their properties of preserving convexity and shap , J. of Zhejiang University (in Chinese).

[6] De Boor, B..(1988), A Practical Guide to Spline, Springer-Verlag, 1988

[7] Wang shefu, Spline Functions and its Applications(in Chinese), West-North Technology

[8] MeiXiangming ,and HuangJingzhi .(1998),Differential Geometry, Higher Education Press

[9] G..E. Farin. Curves and Surfaces for Computer Aided Geometric Design, A practical guide. Academic Press, 1993

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 262-265

E-government Framework Based on Life Cycle of Digital Information Resources

Yan Gao[1, 2]
[1] School of Information Management, Wuhan University, Wuhan, China
[2] School of Economics and Management, Henan University of Science and Technology, Luoyang, China
Email: xingaozhiyuan05@yahoo.com.cn

***Abstract*—This paper proposes a framework for E-government based on information life cycle theory. By literature review, it reveals that it is urgent to establish an integrated E-government framework considering life cycle of digital information resources as an integrated process. Integrating information life cycle theory, the paper provides a framework for E-government, the implications of such a framework would be to inform designers and researchers of E-government about solving problems on domestic digital information development and management, and fulfilling digital information resources management.**

***Index Terms*—e-government, digital information resources, life cycle, framework**

I. INTRODUCTION

Worldwide, local and national government agencies are facing the challenging era of electronic government. A survey of government finance officers reveals that E-government is one of their top concerns (Bornstein, 2000).

While there seems to be substantial growth in the development of E-government initiatives (Bednarz, 2002; Friel, 2002), research on establishing a holistic framework for E-government should always be the priority of government and related agencies. This paper proposes an E-government framework based on information life cycle theory. By literature review, however, it reveals that research on E-government is lack of a holistic framework on theoretical and practical purpose. Therefore, the paper establishes a framework by integrating information life cycle theory. The framework can be utilized to process and analyze other complex issues. The framework can be used to facilitate decision-making unique to each stage and constituency of E-government.

II. LITERATURE REVIEW

Various frameworks of E-government implementation have been advanced in the literature. Three of interest are those proposed by Balutis (2001), Layne and Lee (2001) and the Gartner Group as presented in Baum and DiMaio (2001). These frameworks are detailed in Balutis (2001a) studied 1,300 government agencies and concluded that 57 percent disseminate information and 34 percent allow transactional activity while only 4 percent of the E-government initiatives are "transforming government." The North Carolina Information Resources Management Commission (2001), in a report to that state's General Assembly, looked more deeply into the implications of the Gartner Group framework for practice. Major challenges involved in web presence strategy (phase 1) are: content management, presentation hierarchy, and roles and responsibilities of backend support. Challenges faced as part of interaction strategy (phase 2) are availability of technical support staff and public records management in creating and maintaining databases. One observes the increasing complexity and required investment in technology as we move to phases 2 and 3 (transaction strategy) as the challenges become more complex. These challenges are related to privacy and security, backup and recovery, and internal integration. Even more complex is the advance to the transformation strategy (phase 4).

Other frameworks of E-government have appeared in the literature; however, they are generally descriptive in nature. From these frameworks, some basic propositions for the successful development of E-government have been posited. While this is valuable work at the infancy of E-government, we argue that for E-government to progress a more strategic framework is required. Further, the majority of E-government framework propose a sort of linear progression as E-government evolves, generally beginning with dissemination, then transactions, and finally to some form of integration. We believe that E-government services need not necessarily follow this path. In fact, some may achieve their strategic purpose at the dissemination stage and need not go any further. Since the majority of framework is based on existing E-government applications, which admittedly have been developed on a piecemeal basic, little thought has been given to the development of a coherent strategic portfolio of applications. A framework that begins to broach this topic is sorely needed at this point in the development of the literature.

Other frameworks have been undertaken from application perspective, and they are valuable at the beginning stage of E-government. While our research focuses on the digital information resources of E-government from life cycle perspective to see the forest for the trees, not like other frameworks, the framework not only builds theoretical foundation on E-government, and it also may give practitioners some guidance for further development.

III. E- GOVERNMENT FRAMEWORK BASED ON LIFE CYCLE OF DIGITAL INFORMATION RESOURCES(DIR)

AP-PROC-CS-09CN005

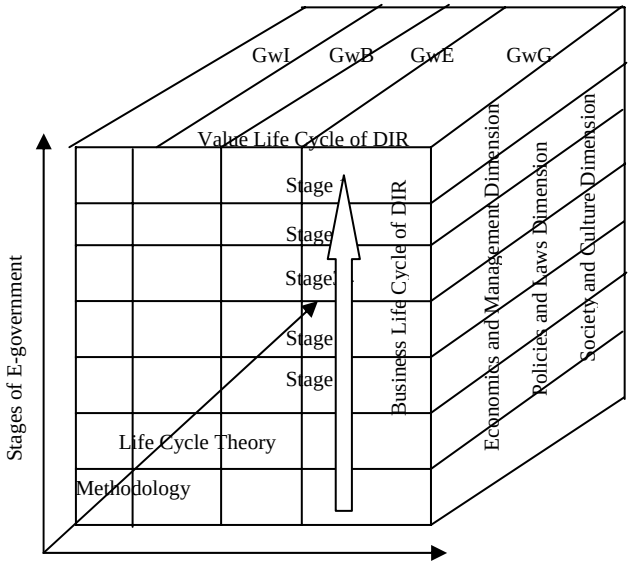


Figure 1. Multi-dimensions framework for E-government based on life cycle of DIR

A multi-dimensions framework for E-government based on life cycle of DIR is proposed in this paper(Figure 1).

A Stages of E-government

X axis stands for business process perspective on DIR.

The stages are presented below, beginning with the least and moving to the most advanced stage of E-government implementation.

1) Information

Information dissemination is the simplest form of E-government where governments post information on web sites for constituents. Thousands of such sites exist. The biggest challenge with these sites, however, is to ensure that the information is available, accurate and timely (Gartner Group, 2000a, b, c). Examples include the US White House informational web page (www.whitehouse.gov/) or the European Union central page (www.europa.eu.int/).

2) Two-way communication

In this stage, government sites allow constituents to communicate with the government and make simple requests and changes. Several of these sites are based on e-mail exchanges, and there are thousand of those as well. Agencies allowing online requests provide sites with fill-in forms but the information is not returned immediately online. It is sent by regular mail or e-mail. An example of this is the US' Social Security Administration web site where constituents can apply for new medicare cards or request benefit statements (www.ssa.gov/).

3) Transaction

At this stage, governments have sites available for actual transactions with constituents. Individuals interact and conduct transactions with the government completely online, whereas these web-based self-services used to be performed by public servants. Actual online transacting is the most sophisticated level of E-government currently widely available. There are several hundreds of these sites. Examples include renewing licenses, paying fines, and applying for financial aid. Benefits of such sites can be very large. For example, the State of Arizona's system to renew vehicle registration online has dramatically reduced waiting lines at department of motor vehicle offices (Thibodeau, 2000).

4) Integration

In this stage, all government services are integrated. This can be accomplished with a single portal that constituents can use to access services they need no matter which agencies or departments offer them. One of the biggest obstacles to more online transactions between the government and its constituents is the lack of integration of all online and back-office systems. Government agencies spend expensive and time-consuming resources to have face-to-face interactions with individuals. For example, in the Kentucky Governor's office up to 90 percent of customer interactions are face-to-face (Thibodeau, 2000). Integrating online systems and back-end systems to support these customer requests could save time and money for the agencies involved, as well as improve customer service. Examples of national portals include the US' FirstGov (http://firstgov.gov/), Australia's State of Victoria's MAXI system (www.maxi.com.au/), and Singapore's eCitizen Centre (www.ecitizen. gov.sg/index_low.html).

5) Participation

These are government sites that provide voting online, registration online or posting comments online. Although this could be seen as a subset of the two-way communication stage, it is so significant as to warrant a separate category. It is helpful to view this as distinct because of the unique sensitivity of providing this online feature. There are few government sites that provide for this level of electronic sophistication.

One of the most prominent future uses of E-government with government may be for individuals to

vote over the internet. A California Internet Voting Task Force reported in 2000 that this must use a phased in approach with great care for authentication and security. Online voting will require technologies to support the privacy of individual voters while allowing re-counts and authentification of identity (Table1).

Table 1. Stages of E-government

	Stages of E-government				
	Stage 1	*Stage 2*	*Stage 3*	*Stage 4*	*Stage 5*
Type of government	Information	Two-way Communication	Transaction	Integration	Political Participation
Government with Individual-Services	Description of Medical benefits	Request& receive individual benefit information	Pay taxes online	All services and entitlements	Not Available (N/A)
Government with Individual-Political Process	Dates of elections	Receive election forms	Receive election funds & disbursement	Register to vote-, province& local (file)	Voting online
Government with Business-Citizen	Regulations online	SEC filings	Pay taxes on line; Apply for and receive program funds Agricultural allotments	All regulatory information on one site	Filing comments online
Government with Business-Marketplace	Posting RFPs (Request for Proposal)	Request clarifications or specs	Online vouchers payments	Marketplace for vendors	N/A
Government with Employees	Pay dates and holiday information	Requests for employment benefit statements	Electronic Paycheck	One-stop shop for info. on job, retirement, vacation, etc.	N/A
Government with Government	Agency filing requirements	Requests from Local governments	Electronic funds transfers		N/A

B Multi-subjects of Development and Management

Y axis stands for various E-government types. Our proposed categorization of E-government types includes:

Government with individuals – delivering services (GwIS). The government establishes or maintains a direct relationship with citizens to deliver a service or benefit. An example is the US' Social Security Administration in its delivery of benefits. This can involve two-way communications as individuals request information about benefits, and government may need information in order to process benefits.

Government with individuals – political process (GwIP). This is the relationship between the government and its citizens as part of the democratic process. It is perhaps the most essential relationship between a government and any entity. Examples include voting online, and participating in requests for comments online during the regulatory process.

Government with business as a citizen (GwBC). Although businesses will not vote, and thus the relationship between businesses and the government will not look exactly like the GwIP, there are still opportunities for business to relate to the government in a citizen-like capacity. Providing securities exchange commission filings online in the US, and paying taxes online in several countries worldwide would be examples of the relationship between government and businesses in this category.

Government with business in the marketplace (GwBMKT). While businesses can receive many online services from government, a major portion of online transactions between governments and businesses involve procurement, or the hiring of contractors or acquisition of goods and services by the government. E-procurement "is one of the fastest growing areas of e-business because it can save time and money" (Symonds, 2000). Some savings reported include 70 percent more efficiency at Australia's Department of Natural Resources and Environment's purchasing department by deploying a paperless system (Symonds, 2000).

Government with employees (GwE). Online relationships between government agencies and their employees face the same requirements as that of the relationships between businesses and their employees. For example, an intranet can be used to provide information to employees, or online transactions with their employees can be performed if agencies have the proper technological architectures. This relationship should be distinguished from the same individual's relationship under GwIP and GwIS.

Government with government (GwG). Government agencies must often collaborate and/or provide services to one another. There are substantial gains from conducting some of these transactions online, between federal, state and local agencies. An example of an inter-governmental level E-government application is the US National Science Foundation's online funding request system called FastLane (www.nsf.gov). The potential for GwG to benefit agencies involved is tremendous – there are over 20,000 web sites for the US Federal Government alone (Thibodeau, 2000).

C Methodology of Framework for E-government

Z axis stands for life cycle theory of digital information resources which is methodology of framework for E-government.

In 1985, Marchand and Horton formally proposed the concept of "information life cycle management", and information management was considered as logical connected several stages or steps, and each step depends on the former one, and the six parts are creating, collecting, organizing, developing, using, cleaning, which illustrated clearly the connectible between information and life cycle. Afterwards, Shenton , Cheatham, Cummins Jenks responded Horton's study by providing an organization view of information life cycle.

In 1994, Hernon argued that the process of public life cycle includes several stages such as creating, collecting, producing, processing, demonstrating, delivering, retrieving and using of information, storage, allocation and display through taxonomic study on earlier discussion of US government information management and life cycle model, and for the first time it analyzed information resources process under electronic context by applying generalized life cycle theory.

In 2000, US "A-130" bill defined "information life cycle" as "the stages of information, and several most important stages are producing or collecting, processing, demonstrating, using, storing and cleaning". Library and Archives of Canada divided it into seven steps, and United States National Committee on Library and Information Science divided into 13 steps.

In 2005, "LIFE" program, launched by British Library and London Metropolitan University, aimed to research on life cycle theory of digital information resources and its application on cost control and investment management, and proposed and established a more complete and standardized life cycle model of digital information resources, including collecting, intaking, metadata, acquiring, storing and preserving.

Based on foresaid researches, the paper proposed business process perspective on digital information resources. The process includes seven stage or steps, they are archive, service, storage, process, collection, distribution and generation.

D Multi-Dimensions of E-government

E-government framework can be studies in four dimensions; they are information technologies, economics and management, policies and laws dimension and society and culture dimension. Information technologies dimension focuses on studying application of information technologies on E-government, such as information systems, data mining, information retrieval. Economics and management dimension focuses on studying improving E-government development economically and strategically. The related resources are information economics, game theory, digital Information resources. Policies and laws dimension are regulations E-government should follow which ensure the rational, scientific development of E-government by legal rights and responsibilities. Researches about government policies, information security are in this field. Society and culture dimension are factors should be considered when undertaking E-government researches. Such as minorities and disable people.

Ⅳ. Conclusion

This paper proposes a framework for E-government based on information life cycle. A literature review justifies the components of framework and reveals that the need for a new framework on E-government. Information life cycle as a methodology is applied to establish a framework for E-government to provide reference for practitioners and researchers to solve their problems. However, further research should be focused on empirical study to test and correct the framework.

References

[1] Adrienne Muir and Charles Oppenheim, National Information Policy developments worldwide I: electronic government[J]. Journal of Information Science 2002; 28; 173

[2] Adrienne Muir and Charles Oppenheim , National Information Policy developments worldwide II: universal access - addressing the digital divide[J]. Journal of Information Science 2002; 28; 263

[3] Beagrie, N and Greenstein, D. (1998). A Strategic Policy Framework for Creating and Preserving Digital Collections[M]. British Library Research and Innovation Report 107. London:The British Library.

[4] Beiser, Karl. Only the FAQs: CD-ROM technology 101[J]. Database, 1994(7):105-111.

[5] C. Oppenheim, J. Stenson,Richard M.S. Wilson.Studies on information as an asset I: definitions[J]. Journal of Information Science, 29 (3) 2003, pp. 159–166

[6] Carolyn Larson, Lori Morse, Georgia Baugh, Amy Boykin, et al. Best Free Reference Web Sites[J]. Reference & User Services Quarterly, 2005(9):39-45.

[7] Cass, Kimberly. Expert systems as general-use advisory tools: An examination of moral responsibility[J]. Business & Professional Ethics Journal, 1996(4):61-86.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 266-269

A Single–depot Complex Vehicle Routing Problem and its PSO Solution

Lei Yin[1], and Xiaoxiang Liu[2]
[1] School of Mechano-Electronic Engineering Xidian University, Xi'an, China
Email: yinlei_w@163.com
[2] Department of Computer Science of Zhuhai College Jinan University, Zhuhai, China
Email: tlxx@jnu.edu.cn

***Abstract*—This paper discusses a single-depot complex vehicle routing problem (SCVRP), of which the conditions are that the total routine shall be the shortest and the biggest marched routine of any single vehicle as well. A single objective model is set up upon these conditions. Accordingly, this paper proposes an improved PSO algorithm (GPSO) combined with the crossover operation of genetic algorithm (GA). It can avoid being trapped in local optimum due to using probability searching. GPSO is applied to SCVRP, and cases testify to its feasibility and effectiveness.**

***Index Terms*—Particle Swarm Optimization, vehicle routing problem, genetic algorithm**

I. INTRODUCTION

Vehicle Routing Problem (VRP) was first proposed by Dantzig and Ramser in 1959. It means to design an adequate driving routing for a series of dispatching sites so that the vehicles can go through these sites orderly, and under certain constraint conditions achieves some objectives (such as covering the shortest route, costing the least, consuming the least amount of time etc.) [1]. This problem is a complete NP problem and of great importance in disciplines like operations research, computer science, logistics and administration. At present the solutions of VRP can mainly be divided into three categories [2]: the simple heuristic algorithm, the mixed algorithm of heuristic constraint planning and local search, accurate optimization algorithm and intelligent optimization algorithm, such as genetic algorithm, taboo search, stimulated annealing algorithm and improved particle swarm algorithm.

Numerous scholars are doing research on VRP till now. They touch upon various methods, while most of which are for comparatively simple VRP problems [3]. To apply the lab research achievements to real logistic delivery process, there is still a large amount of work to do. Probable factors like vehicle type, time window, capacity constraint, road conditions, and weather are to be considered, which is very difficult.. Therefore, traditional methods can hardly help to find the optimized route [4-5].

Particle swarm optimization [6] (PSO) is an improved computing technology, as well as a bionic algorithm which simulates the flight of bird flock with advantages of few individual numbers, easy computing and nice robustness. This paper puts forward the GPSO, puts it into the solving of single-depot complex vehicle routing problems and has obtained ideal effects.

II. DESCRIPTION OF GENERAL VEHICLE ROUTING PROBLEMS AND THE MATHEMATICAL MODELING

General vehicle routing problems can be described as: one central warehouse, number of vehicles K, respective capacities q_k (k=1, 2,…, K); and there are L dispatching tasks, represented as 1,2,…,L. The freight volume of the dispatching site i is g_i (i=1, 2,…, L). The goal is to seek the vehicle routing that costs the least while meeting the freight requirements.

Ref. [1] numbers the central warehouse as 0, the dispatching sites 1, 2,…, L, and tasks and the central warehouse are represented as i (i=0, 1,…, L). Variables are defined as:

$$y_{ki} = \begin{cases} 1, \text{the task } i \text{ is completed by vehicle } k \\ 0, \text{otherwise} \end{cases},$$

$$x_{ijk} = \begin{cases} 1, \text{vehicle } k \text{ travels from point } i \text{ to point } j \\ 0, \text{otherwise} \end{cases},$$

c_{ij} represents the transportation cost from i to j, referring to distance, cost or time etc.

Thus the mathematical modeling for vehicle optimization dispatching can be obtained as follows:

$$\min z = \sum_i \sum_j \sum_k c_{ij} x_{ijk} + p_E \sum_{i=1}^{L} \max(ET_i - s_i, 0) + p_L \sum_{i=1}^{L} \max(s_i - ET_i, 0) \quad (1)$$

$$\text{s.t.} \quad \sum_i g_i y_{ki} \le q_k \quad \forall k \quad (2)$$

$$\sum_k y_{ki} = 1 \quad i = 1,2,\cdots,L \quad (3)$$

$$\sum_i x_{ijk} = y_{kj} \quad j = 0,1,\cdots,L; \ \forall k \quad (4)$$

$$\sum_j x_{ijk} = y_{ki} \quad i = 0,1,\cdots,L; \ \forall k \quad (5)$$

$$x_{ijk}, y_{ki} = 0 \text{or} 1 \quad i, j = 0,1,\cdots,L; \ \forall k \quad (6)$$

corresponding author: tlxx@jnu.edu.cn

III. A SINGLE-DEPOT COMPLEX VEHICLE ROUTING PROBLEM (SCVRP)

A. Description of the problem

Single-depot vehicle routing problems are based on general vehicle routing problems as mentioned previously. Only some conditions make the problems more complex.

1) The plant area has many sites. The freight tasks between them are undertaken by a number of indistinctive vehicles;

2) In one dispatching task, vehicles set off from the fixed departure center and have to return back in the end;

3) The freight tasks between sites may exceed the capacity of the vehicles;

4) There is limited time for one single dispatching task, which means, all the vehicles shall return within regulated time.

B. Assumptions about the problem and objectives

First here are some assumptions about the problem:

1) Suppose the vehicle speed is invariable, uninfluenced by the dispatching plans. Reasons for the assumption: the road conditions of the plant area are satisfying and there is a speed limit out of the concern about safety. The changing of the vehicle routing is not the main cause for speed variation.

2) We suppose that there is no time window constraint on the routing. Reason for the assumption: the goods to be transported in the plant area are the products of the previous working cycle. What the goods need is only to be delivered in the current cycle.

According to the above problem description and assumptions, two objectives are concluded.

Objective 1: To make the total routing the shortest. In this problem, transportation of goods does not bring the plant direct economic benefits, i.e. transportation does not produce value. Contrarily, transportation will cost a tremendous sum of money. Thus to cut down on transportation cost to the utmost is desired. When vehicles and staff are fixed, the vehicle routing decides the cost of transportation. Regarding economy, the plant may take the shortest total routine as an objective.

Objective 2: To make the biggest marched routing of any single vehicle the shortest. Shortest total routing doubtlessly reduce the cost for the plant, but some vehicles probably have to transport far more than others for that objective. A series of management problems may turn up, like the staff become unsatisfied owing to the difference of work loads, vehicles are unable to return to the departure center, leading to handover disorder, and such. So regarding management, the plant must consider **objective 2**.

Objective 1 and **objective 2** are both of great realistic significance. However, they are contradictory to each other. They may be considered differently according to real situations. This paper takes **objective 1** as the final objective, while **objective 2** is regarded as a constraint (That is, the biggest marched routine of a single vehicle cannot exceed a certain number).

IV. AN IMPROVED PSO FOR SOLVING SINGLE-DEPOT COMPLEX VEHICLE ROUTING PROBLEMS

In general PSO algorithms, the speed of particles v are limited by the maximum speed v_{max}. It determines the searching precision of the solution space of particles. If v_{max} is too high, particles may exceed the optimal solution; if v_{max} is too low, particles may fall into local search space while missing the global search [7]. So this paper resorts to the crossover operation of genetic algorithm, embraces the advantages of particle swarm optimization algorithm, and puts forward an improved particle swarm optimization algorithm (GPSO) aimed at solving single depot complex vehicle routine problems, as elaborated below.

A. Coding and Decoding

Firstly, sites and tasks need to be coded. For instance, the departure center is number as 0, sites *A*, *B*, *C* respectively 1, 2, 3. In one dispatching run, vehicles need to undertake a series of tasks. Tasks from *A* to *B* (*A*→*B*), from *B* to *C* (*B*→*C*) are numbered as 1, 2.

The problem is described as: there are n tasks and m vehicles. Each task needs to be performed Pi times (i=1, 2,…, n). Thus the coding length is $n\times Pi+m$-1. Of them, the anterior $n\times Pi$ are task positions, the posterior m-1 are decoding positions. In task positions, the times that tasks appear should be in accord with Pi.

More elaboration below:

Suppose there are 2 tasks, separately need to be undertaken for 2 times and 3 times. There are 2 vehicles for dispatching. And there is 1 particle X=[1 1 2 2 1 2 4]. Then [1 1 2 2 1] are task positions, [2 4] are decoding positions.

The order that corresponding vehicles of the particle perform the tasks is:

Vehicle 1: 0→1→1→0;

Vehicle 2: 0→2→2→0;

Vehicle 3: 0→1→0.

Because task 1 represents (*A*→*B*), task 2 represents (*B*→*C*), and the number of *A*, *B*, *C* are 1, 2, 3, in final, the corresponding site routing of the particle is:

Vehicle 1: 0→1→2→1→2→0;

Vehicle 2: 0→2→3→2→3→0;

Vehicle 3: 0→1→2→0.

B. Crossover Operation

In order to preserve the good gene fragments of parents, this paper proposes a crossover operation which is inspired by genetic algorithm. Suppose two parent individuals are f_1 and f_2, through crossover operation to get the offspring individual *new*. Specific operations are as follows:

1) Remove the decoding positions of f_1 and f_2, get new strings f_1' and f_2'. The length of strings is $n\times P_i$, in which the requisite task sequence is *A*.

2) Randomly choose a cross region (*a*, *b*) in the second string.

3) Add the numbers in the cross region of f_2' to the front of f_1', delete the numbers of the cross region of f_2' in f_1' and get a sub-string new_1. Add the numbers in the

cross region of f_1' to the front of f_2', delete the numbers of the cross region of f_1' in f_2' and get a sub-string new_2.

4) Turn new_1 and new_2 into strings that contain requisite tasks and get new_1' and new_2'. Specifically, add $i/10$ to every position number in new_1 and new_2, and value them from small to big as integers 1 to $n \times P_i$ and get $W1$ and $W2$, thus, $new_{1i}' = A_{W1i}$, $new_{2i}' = A_{W2i}$

5) Renewedly produce decoding positions randomly.

6) Calculate adaptive values with (1), retain the offspring with a lower adaptive value recorded as *new*.

If the two parent individuals are:

f_1=1 2 3 4 5 6 1 2 1 3 7, f_2=1 1 2 6 5 4 3 2 1 3 7 cross region is: (3, 7), requisite task included is: A=1 1 1 2 2 3 4 5 6. Then f_1'=1 2 3 4 5 6 1 2 1, f_2'=1 1 2 6 5 4 3 2 1. After crossover operation new_1=6 5 4 3 1 2 3 2 1, new_2=4 5 6 1 1 1 2 2 1. $W1$=9 8 7 5 1 3 6 4 2, $W2$=7 8 9 1 2 3 5 6 4, new_1'=6 5 4 2 1 1 3 2 1, new_2'=4 5 6 1 1 1 2 3 2.

Calculate the adaptive values of new_1 and new_2 with (1), and take the string with the lowest adaptive value as *new*. This way, the crossover operation between two parent strings is finished, and we get the offspring generation *new*.

C. Algorithm

In this paper, we use the hybrid PSO to solve SCVRP, whose pseudocode of algorithm is described in Fig. 1.

V. EXPERIMENT AND RESULT

A plant area has departure center *A*, 3 distinctive vehicles, 4 operation sites *A*, *B*, *C*, *D* (Numbering see Table 1). Now there are 5 transportation tasks (Numbered as 1, 2,…, 5). The frequencies of transportation of each task as Tab. 1. The distances between departure center *A* and operation sites d_{ij} as Tab. 2. It is required that the biggest marched routing of any single vehicle *z* cannot exceed 22.

```
Set iteration times for MaxN, randomly generated initial particles N;
Calculate the fitness of initial particle to be l0. According to the initial fitness of each particle, initialize plbesti, pxbesti, glbest and gxbest; L represents objective 2; its threshold value is E.
WHILE (iteration times< MaxN) DO
  FOR i=1: MaxN
    Z=1;
    While (Z)
      The ith particle Xi crosses with gxbest and selection to be Xi'; Xi' crosses with pxbesti and selection to be Xi'', ordering Xi=Xi'';
      If L<E
        Z=0;
      End if
    End while
    Calculate fitness li according to current location;
    IF (li< plbesti)
      pxbesti=Xi, plbesti=li;
    END IF
  END FOR
  Update plbesti, pxbesti, glbest, gxbest;
END WHILE
Finally, export glbest and gxbest.
```

Figure 1. Pseudocode of GPSO algorithm to solve SCVRP

TABLE I. FEATURES AND REQUIREMENTS OF TASKS

Number	Site	Number	Task	Frequency
1	*A*	1	*A*→*B*	3
2	*B*	2	*A*→*C*	2
3	*C*	3	*C*→*B*	2
4	*D*	4	*D*→*A*	1
		5	*C*→*D*	2

TABLE II. DISTANCES BETWEEN SITES

d_{ij}	1	2	3	4
1	0	3	6	5
2	3	0	3	4
3	6	3	0	5
4	5	4	5	0

The known optimal result is: the minimum total marched routine Z=62, the minimum of the biggest marched routines of any single vehicle z=30. Coincidentally, the two can both be achieved in this case. As for the conditions of this case, they can hardly satisfy to meet $z \leq 30$. Experiments are performed under different parameter settings so as to research on the functions of the improved PSO mentioned in the paper.(Experiment environment: 1GB, Inter P4) Results see Tab.3 below.

When n=70, N=70, the success rate of algorithm search reaches 100%, the operation frequency is 70×70=4900, operation time is 9.35s (Under this parameter setting, operation runs 100 times, the average iteration of Z is as in Fig.2 below). While if the same problem is solved with enumeration method, the operation frequency will be:

$$A_x^{\ x} C_{x-1}^{\ y-1} = A_{10}^{\ 10} C_9^{\ 2} = 130636800$$

x is the number of sites, y is the number of vehicles, operation time is about 3 days in the same experiment environment. Obviously, GPSO can better meet the requirements of practical projects.

TABLE III. IMPROVED PSO SOLVES 100 SEEKS THE NUMBER OF THE OPTIMAL SOLUTION NO. IN DIFFERENT PARTICLE NUMBERS *N* AND DIFFERENT ITERATION NUMBERS *N*

NO.	*n*=50	*n*=60	*n*=70	*n*=80	*n*=90
N=50	85	88	90	93	97
N=60	89	92	94	98	100
N=70	93	94	100	100	100
N=80	97	98	100	100	100
N=90	98	99	100	100	100

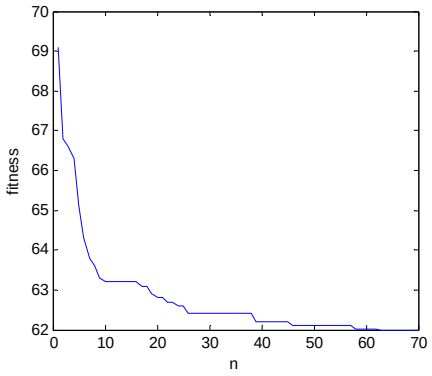


Figure 2. The algorithm convergence graph of the improved PSO when n=70, N=70

REFERENCES

[1] Li Jun and Guo Yaohuang, "Dispatch optimize theory and methods for logistic distribution vehicle," Beijing:Chinese Commodity Publish House, 2001.

[2] Hoong C L, Melvyn S, and Kwong M T, "Vehicle routing problem with time windows and a limited number of vehicles," European Journal of Operational Re-search, vol. 148, pp. 559–569, 2003.

[3] Zhong Shiquan and He Guoguang, "Vehicle Scheduling Problem with Single Depot in Complex Conditions," Systems Engineering, vol. 23, pp. 29–32, 2005.

[4] Cao Hongmei, Gao Li, and Hu Yaxin, "Hybrid Algorithm for Solving Variable Fleet Vehicle Routing Problem," Journal of Wuhan University of Technology, vol. 33, pp. 647–650, 2009.

[5] Liao Liangcai, Wang Dong, and Zhou Feng, "Solving Method of the Optimization Problem of Logistic Distribution Vehicle Scheduling Based on Hybrid Genetic Algorithm," Systems Engineering, vol. 26, pp. 27–31, 2008.

[6] R C Eberhart and J Kennedy, "A New Optimizer Using Particles Swarm Theory," In: Proc Sixth International Symposium on Micro Machine and Human Science, Nagoya, Japan, 1995.

[7] Ding Haili, Wang Fang, and Gao Chengxiu, "Crossover particle swarm optimization for traveling salesman problem," Journal of Mathematics, vol. 8, pp. 85–89, 2008.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 270-273

New VPN Application in 3G Network

Weili Huang[1], and Jian Yang[2]
[1] Hebei University of Engineering , Handan, Hebei 056038, China
997518268@qq.com
[2] Hebei University of Engineering ,Handan, Hebei 056038, China
yjbs11@126.com

Abstract—**This article will use VPN technology for new wireless 3G networks, We propose the new architecture and mechanisms of VPN in the 3G.With the new view Parallel Server Cluster and MPLS-VPN-based algorithm and the principles of VPN technology, we discover the current vulnerability of VPN technology in the 3G network and prospect its application in the 3G network. We are contacting cloud computing and VPN to study cloud security with VPN. This will be closely integrated VPN security with the clouds, to discover the new applications of VPN security in the cloud computing .**

***Index Terms*—VPN;3G;Cloud Computing; Parallel Cluste; New Structure; Constraint-based routing**

I. INTRODUCTION

With the advent of 3G networks, a variety of technologies to meet the 3G network will be mature. 3G network is high-speed ,wireless and mobility,providing a fiber optic line, ADSL broadband access can not match the convenience,so it is gradually becoming indispensable to the current broadband access in a complementary manner. In some non-fixed location, cable broadband can not reach there, but they require high-bandwidth access to the environment play a unique role. For such a nascent Internet age, the network speed will remain the focus of attention. We had a lot of trial operation to increase network access speed, with mixed success. However, presentation and application of VPN (Virtual Private Network)in the 3G network has brought the gospel. In fact, cable network VPN has existed for long time, but accessing to 3G networks later, instead of the VPN routing, VPN hardware firewalls, etc. are based on the radio and there. Except wireless Internet, in the application layer,because passing through the operator's public network, so security must be taken into account, and because the tense of the current global IPV4 address, generally the ip address from 3G network can not directly access the internet ,but VPN can create their own virtual local area network business, just from the security and accessibility are two aspects of a good solution to both problems. In recent years, VPN technology has been widely used .For business, VPN's biggest attraction is price. It is estimated that, if an enterprise abandon the leased-line and use VPN, the cost of their entire network can save 21% -45% .Thus, VPN in the 3G network have great commercial prospects, therefore, there are a lot of third-party vendor for VPN[1].

II. CLOUD COMPUTING AND VPN

Recently,the cloud security has become the popular network-based security technology for a lot of network providers .Cloud security is an effective security model, and soon will become the mainstream. But, it has not been accepted by the majority of users, the reasons is that cloud providers can not guarantee the security of the technology itself. Because of the special structure of cloud technology, it had a lot of big loopholes.So people put a VPN security technology into the cloud technology, we need to be established among the clouds. The users could control security for independent cloud, crossing cloud and multi-cloud.

Controling information ,in fact,not rely on a fixed position. A simple example is the public key cryptography. I insist the ownership of the private key so that I can control the user. Usually, the private key is stored in a secure location.However, from the ownership of the key I can try to control information, and not necessarily own the ownership for other infrastructures. I can create a trusted VPN by a incredible infrastructure . We can control and protect the security of information by a key connection ,and can connect service agreement[2].

If these are very in place, then there is no inherent reason to make cloud computing environment can not be guaranteed safe. In order to ensure safety we do not maintain these things. We can establish a trusted VPN. Safety audit staff and mediators constant introduction of new technologies and business models. If we can clearly show that we can control security by technology and connectivity, we should make a cloud computing environment safe and reliable as a private facility.

However, we can also use the "multi-VPN technology" to achieve multiple security detection . The so-called "multi-VPN technology", that is, to set up multiple VPN channels between two points.Before an accessing point achieve another accessing point ,it must be crossed multi-layer filter,in order to achieve further improving security . Fig.1:

III. MPLS-BASED VPN TECHNOLOGY IN THE 3G NETWORK

A. VPN Routing

Traditional VPN transmited private network data flow In the public Internet with GRE, L2TP, PPTP tunneling protocol, LSP tunnel itself is in a public Internet. So, to achieve the VPN using MPLS has a natural advantage. MPLS VPN is a private network through the LSP to the different branches of banded together to form a unified

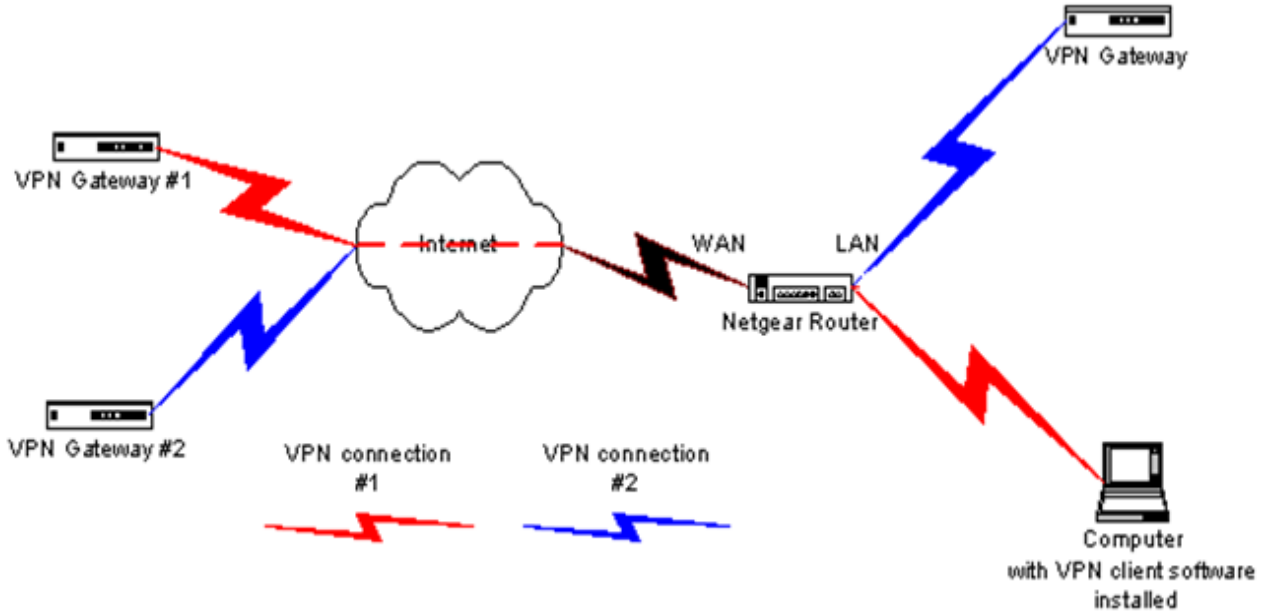


Figure.1 Multi-VPN

network. MPLS VPN also supports interoperability between different VPN control. MPLS VPN supports the reuse of IP addresses among different branches, and support interoperability between different VPN. Compared with traditional routing, VPN routing needs to increase the branching and VPN-identifying information, which needs to extend BGP protocol to carry VPN routing informations.

B. *Constraint-Based Routing Calculation*

For the data stream of constructing transmission path, if it had no the specific transmission requirements, we can follow the traditional routing method(the shortest path) to establish transmission path . But if the data stream had a clear demand for services, you need to follow the service flow and the actual state of the network to obtain a suitable transmission path.Constraint-based routing technology is a good solution, it can be used to calculate many routes based on a variety of constraints, such as under the data flow requirements, available network resources situation and the strategies of the network administrator we can calculate a comfortable path, which not only to meet the data flow requirements but also to focus on optimizing network resource utilization.We can see that, compared with traditional constraint-based routing calculation which was only considered network topology, constraint-based routing technology must consider the network topology, the distribution of network resources situation, the administrator's strategy and business flow requirement. So constraint-based routing are able to find a link may be long but light load, rather than a heavy load of the shortest path, so that the network load distribution becomes uniform, to avoid network transmission in the hot spot . The focal points of constraint-based routing can be summarized as follows[3]:

(1) Screening the network link set which was in line with data flow requirement and the administrator's strategies.

(2) Using the shortest path algorithm on the rest of the topology map.

IV. NEW VPN APPLICATION IN THE 3G NETWORK

A. *.The 3G Network-based VPN Access to Image Data*

VPN is based on the existing network to establish a virtual LAN, which means the equipment in the network or server has two ip addresses, one is pre-wired ip address, it is wired ,such as the server's ip address 212.25.4.1 is public network IP addresses, the scope of VPN network segment address we program is 192.168.2.1-255, assuming that we assign to the VPN server, the ip address of 192.168.2.1, when a device-side firstly get an ip address 118.34 .2.15 by 3G wireless network, but the ip address can not take the initiative to access from the outside , and then the device have an establishment of the VPN server's virtual connection by the built-in PPTP VPN client. It is like the device have two network cards and establish direct connection to the network server.The VPN server will assign the virtual links an ip address of 192.168.2.2 , so that it can visit the device 192.168.2.2 by using 192.168.2.1 .At this time if we need a remote watching for an image data from the device , first we have a computer which could access to the Internet , and then have an establishment of the VPN server,that is to say , the client get the ip address of 192.168.2.3 which is from the VPN server , so that the client and the central VPN server and the device just like in a real LAN, you can access each other. Clients can monitor remote image with the vendor-supplied client software or the ip address 192.168.2.2 in the IE browser[4-6].

B. *New VPN Architecture In 3G Network*

- The current most prominent feature of 3G networks are not subject to geographical constraints, the stability and the speed of wireless transmission are the most important indicator. First, in the wireless router , we need to have a powerful processing chip and its function should be a fast process data and fast

forwarding, for example, I recommend TI davinci chip, the overall use of “ARM + DSP ”architecture,.Namely, it associated the use of ARM Embedded Technology with DSP-processing structure.

- Second, there must be the VPN-based server. Because the ip address of wireless Internet is not the real ip address which can be routed on the internet, so the client can not link point to point with video servers and you must transit with the central VPN server.
- The third is to build a virtual network tunnels. In the VPN, using PPP (Point to Point Protocol) packet stream is from a router on the LAN issue, through a shared IP network to transmit encrypted tunnel, and then to another router on the LAN, so that The tunnel is instead of dedicated lines[7]..Fig.2:

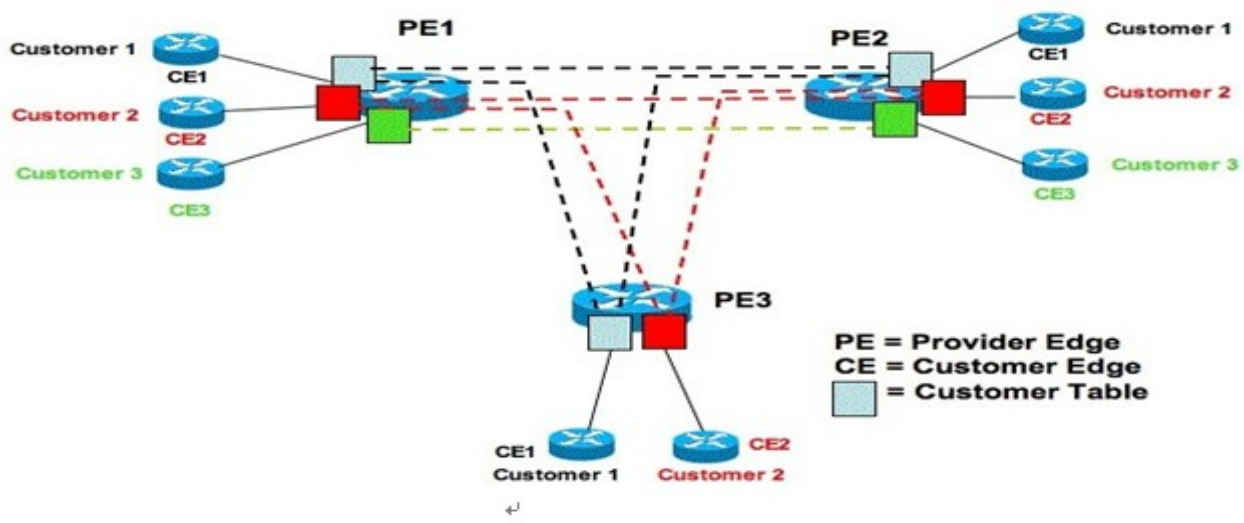


Figure.2 New VPN Architecture

C. New VPN Mechanism In 3G Networ

3G network has just been put into practice, to some extent,a variety of mechanisms are not perfect. On the mainland ,most of PPPoE dial-up mechanism was used to access to the network environment, the operators change IP address frequently, which led to the instability of the network data transmission, however, in the 3G network, VPN have some new mechanisms to ensure the stability of its data transmission, they are: DPD (Dead Peer Detection) to detect VPN disconnection mechanism and to clear to see that whether online, in order to ensure that the VPN disconnected, it can make a first response and management; Keep-Alive mechanism, to continue to maintain a VPN line to help businesses automatically attempt to reach the first line of VPN to work to further ensure the VPN the stability of the quality of services are not dropped; NATT (NAT Traversal), ensure that the VPN device compatible with the mechanism, which is the mechanism for converting the packet format can be issued by the ESP Enterprise IPSec packet format into the UDP format, and then through the office management the Center's core router, VPN information can flow to achieve the purpose; VPN guaranteed bandwidth mechanism; DDNS redundant VPN increase stability mechanism, which can help two dynamic IP-VPN gateways to find each other[8] .

V. PARALLEL VPN SERVER CLUSTERS

Unlike traditional VPN server, to its current position,there are still some gaps between the VPN server in the 3G network and traditional VPN server. However, in the 3G network to improve data transmission speed is what we pursue, which means that operators have a more rapid approach to data forwarding. With the development of the algorithm, we know that in a sense, parallel computing method is superior to a serial operation, but we want the clusters in parallel to further parallel them, will be quicker to
do so? So I have proposed a parallel cluster's claim that on the VPN server-side, firstly,with a service delivery point of an area as the unit,we parallel combination of the server machines so as to achieve rapid data processing purposes , then, with the whole Chinese 3G network as the unit, we parallel combination of very service delivery point of an area so as to reach the data processing and transmission of data synchronization purposes, namely, the whole Chinese 3G network is an entirety, rather than the various regions for their own array. They can get each other's support and use resources.This approach in the traditional network service is not easy to achieve, but the 3G network, it can be easily achieved, because in the 3G network, data transmission is wireless, and the platform they built is wireless, it broke through space and wiring constraints, the idea of the parallel cluster is also relatively easy to achieve[9-12].

Exchange of information and data can still make use of existing network protocols, but I would suggest to make use of the operating system Linux, on the one hand because of Linux to access the network operations more efficient, on the other, the higher their safety and can reduce costs overhead.

In addition, in order to achieve a server cluster in parallel with traditional VPN server, there is a different storage systems, and on security on the requirements.

Storage system must provide the high performance of I / O operations and data for Linux cluster in order to meet a substantial number of server aggregation of demand, of course, each computer must be in parallel computers, in every small 3G cluster node to install a sufficient number of chassis to provide additional capacity, through a sufficient number of Gigabit Ethernet port tied together and multiple cluster nodes connect to the Gigabit Ethernet switch, so that each node and the entire storage space can be the establishment of the relationship between the parallel data access so that it can effectively improve storage system performance. I say "sufficient number " of the specific number depends on the specific circumstances and the amount of data processed, depending mainly storage and unobstructed access to the standard.

We have to parallel the VPN server, we must share database, meaning that the entire 3G network data in a database must be accessible to each other, it is not recommended the establishment of a unified database of 3G, but rather through sharing agreement with each other in all the database access, we have to use the same database format, but the issue of security also will bring us new troubles. Since the implementation of data sharing, then there is large amounts of data may be stolen during transmission, or by others, the possibility of eavesdropping. [13-15]

VPN can form a virtual LAN, and this virtual local area network where the computer is completely mutual trust.The attack among these computers is very difficult to be on guard. So,firstly, the VPN vendors should set up the initial state well, let non-local area network visits. Secondly, we should strengthen the development of firewall, VPN, can not be limited to such as MD5 or AES encryption algorithm common to support the transmission of data, but also should be embedded VPN to a variety of systems so as to achieve full transparency of the VPN infrastructure[16-17].

VI. CONCLUSION

In the emerging 3G networks ,using VPN technology, we can have a good access speed, especially if we join cluster parallel technology in this paper to the VPN server, because this technology include the parallel machine and parallel algorithm,with which effectively the world's 3G networks all nodes are utilized, and even more effectively improve the performance of VPN access to the network itself. In addition, VPN's new architecture and mechanisms put forward, it also breaks the traditional framework of constraints, give full play to the 3G network, wireless connectivity, but in their access to data security issues should also be further to strengthen, should develop a new more secure encryption algorithm, to deal with wireless transmission.

REFERENCES

[1] Haiying Gao, VPN Technology[M],Machinery Industry Press,2004,pp.72-99. (in chinese)

[2] (US) Lucas . Firewall Policy and VPN Configuration . Water Resources and Hydropower Press,2008,pp.321-444.

[3] Zhiying Lv. On the VPN Technology .《Management and Technology》Journals, 2008 No. 3.pp.34-77.

[4] Dengguo Feng . Network Security Principles and Technology. Science Press, 2003.pp.65-87.(in chinese)

[5] Elizabeth D.Zwicky,Simon Cooper,Tsinghua University Press,2003,pp.54-77.

[6] Carasic-Hengmu.Firewall Core Technology Intensive Solution. Hydropower Press. 2005. 4.pp..99-102

[7] Mei Zhang.SSL VPN Key Technology Research and System Design[D].PLA Information Engineering University Press,2006,pp.89-99.

[8] Jiazhen Xu,Comparison and Analysis of IPSec-based and SSL-based VPN[J],Computer Engineering and Design Press 2004,pp.99-105. (in chinese)

[9] Freier,karhon.The SSl Protocol Version 3,Netscape Communications November 18,1996.

[10] David Wagner,Analysis of SSL 3.0 Protocol,Http://www.Counterpane.com,2005-08-07.

[11] Rivest RLShamir A,Adleman LA. Method for obtaining digital signatures and public key cryptosystems. [J]CACM1978~21(2),pp120-122

[12] KaiCheng Lu,. Computer cryptography【M】Beijing, Tsinghua University Press. . 1998. 7. pp73～75 (in chinese)

[13] DaWang.The VPN Interpretation,Netscape Tsinghua University Press November 18,1996. (in chinese)

[14] http://www.baidu.com

[15] Ivan Pepelnjak,Jim Guichard.MPLS and VPN Architectures [M]. Beijing Posts & Telecom Press , 2001pp.89 -109, 129-155

[16] Eric W .Gray.MPLS:Implementing the Technology [M]. Electronics Industry Press , 2003.pp119- 122

[17] Thomas M. Thomas II. OSPF Network Design Solutions Second Edition[M]. Electronics Industry Press, 2004,pp25-42

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 274-277

Application Research of k-means Clustering Algorithm in Image Retrieval System

Hong Liu [1], and Xiaohong Yu [2]
[1] College of Computer science and Information Engineering Zhejiang Gongshang University, HangZhou, China
Email: LLH @mail.hzic.edu.cn
[2] College of Computer science and Information Engineering Zhejiang Gongshang University, HangZhou, China
Email: XHYU @mail.zjgsu.edu.cn

***Abstract*—In image retrieval algorithms, retrieval is according to feature similarities with respect to the query, ignoring the similarities among images in database. To use the feature similarities information, this paper presents an application of k-means clustering algorithm to image retrieval system. Combining the low-level visual features and high-level concepts, the proposed approach fully explores the similarities among images in database, using such clustering algorithm and optimizes the relevance results from traditional image retrieval system by firstly clustering the similar images in the images database to improve the efficiency of images retrieval system. The results of experiments on the testing images show that the proposed approach can greatly improve the efficiency and performances of image retrieval, as well as the convergence to user's retrieval concept.**

***Index Terms*—image retrieval, k-means cluster algorithm, feature extraction**

I. INTRODUCTION

With the popularity of internet and rapid development of digital technique, content-based image retrieval (CBIR) has become an important part of information retrieval technology. CBIR technique focus on searching images in database similar to the query image, according to the image features related to content. This technique is based on automatic extraction of image features, retrieves by automatically comparing the features of query image (such as color, shape, texture, etc.). With the corresponding features in image feature library, and finally outputs the best matching images and its corresponding information.

In CBIR, feature vectors extracted from images usually exist in a very high-dimensional space, such high dimensionality of the feature vectors leads to high computational complexity in calculation for similarity retrieval, and inefficiency in indexing and search, and where a parametric characterization of the distribution is often impossible. Due to the high dimensionality, researchers use the similarity measure to measure the degree of similarity between images. But, there still exists a semantic gap, which just reflects the discrepancy between the relatively limited descriptive power of low-level visual features and high-level concepts. The system is based on the similarities between the query image and images in database while ignoring the similarities among images in database. In order to solve this problem, graph theoretic approaches have been used to effectively explore the similarities among images in database, and the problem could be transformed into solving a maximal clique problem in graph theory which is a clustering problem in computer vision area.

Nowadays, there are also some researches about it. For example, "Texts" in natural languages are the main means to convey semantics among human being. Therefore, the status quo of semantic image clustering is to incorporate ''texts", e.g., captions to facilitate the understanding of images. Gong et al. [5] proposed to integrate the captions of images for semantic clustering; Dai and Cai [6] built a semantic tolerance model which first represents images based on semantic classification. Hai [7] proposes to understand the images through the analysis of semantic links existing among web pages. In this area where the ever-increasing number of images acquired through the digital world, it makes the brute force searching almost impossible. A user's query interest is often focused on one particular part of the image, i.e., a region in the image that has an obvious semantic meaning. Therefore, rather than viewing each image as a whole, it is more reasonable to view it as a set of semantic regions Of course, for such a problem, some people do some research. In [8, 10], it is proposed that semantic clustering is performed using relevance feedback. These works are based on the whole image instead of image sub-regions/regions. The clustering method in [10] is based on a method called CAST while the one in [8] is based on the Association Rule Hyper-graph Partitioning algorithm, etc.

This paper puts forward a new framework of content-based image retrieval, which integrates semantic cluster classifier with k-means algorithm. And to improve the efficiency, we propose to impose a clustering component in the region-based image retrieval, which makes it possible to only search the clusters that are close to the query target, instead of searching the whole search space. The rest of the paper is organized as follows. The algorithms for k-means clustering are introduced in Section 2. The new image retrieval algorithm framework is described in Section 3. Experiments and results are presented in Section 4. Finally, conclusions and future works are given in Section 5.

II. K-MEANS CLUSTERING

Clustering algorithm has been widely used in computer vision such as image segmentation and

database organization. The purpose of clustering is to group images whose feature vectors are similar by similarity judgment standard; meanwhile to separate the dissimilar images. Clustering algorithms can be broadly divided into two groups: hierarchical and partitional. Hierarchical clustering algorithms recursively find nested clusters either in agglomerativemode (starting with each data point in its own cluster and merging the most similar pair of clusters successively to form a cluster hierarchy) or in divisive (top-down) mode (starting with all the data points in one cluster and recursively dividing each cluster into smaller clusters). Compared to hierarchical clustering algorithms, partitional clustering algorithms find all the clusters simultaneously as a partition of the data and do not impose a hierarchical structure. Input to a hierarchical algorithm is an n*n similarity matrix, where n is the number of objects to be clustered. On the other hand, a partitional algorithm can use either an n*d pattern matrix, where n objects are embedded in a d-dimensional feature space, or an n*n similarity matrix. Note that a similarity matrix can be easily derived from a pattern matrix, but ordination methods such as multi-dimensional scaling (MDS) are needed to derive a pattern matrix from a similarity matrix.

The most well-known hierarchical algorithms are single-link and complete-link; the most popular and the simplest partitional algorithm is K-means. Since partitional algorithms are preferred in pattern recognition due to the nature of available data, K-means has a rich and diverse history as it was independently discovered in different scientific fields, it is one of the most widely used algorithms for clustering. Ease of implementation, simplicity, efficiency, and empirical success are the main reasons for its popularity. In the paper, we apply k-means algorithm to analysis images similarities in the database.

Let $X=\{x_i\}$,i=1,…..n be the set of n d-dimensional points to be clustered into a set of k clusters, $C=\{c_k$,k=1,….k}; k-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized. Let u_k be the mean of cluster c_k. The squared error between u_k and the points in cluster c_k is defined as

$$J(c_k) = \sum_{x_i \in c_k} \|x_i - u_k\|^2 . \qquad (1)$$

The goal of K-means is to minimize the sum of the squared error over all K clusters,

$$J(c_k) = \sum_{k=1}^{k} \sum_{x_i \in c_k} \|x_i - u_k\|^2 . \qquad (2)$$

Minimizing this objective function is known to be an NP-hard problem (even for k =2). Thus K-means, which is a greedy algorithm, can only converge to a local minimum, even though recent study has shown with a large probability k-means could converge to the global optimum when clusters are well separated (Meila, 2006). k-means starts with an initial partition with k clusters and assign patterns to clusters so as to reduce the squared error. Since the squared error always decrease with an increase in the number of clusters k (with J(C)= 0 when k= n), it can be minimized only for a fixed number of clusters. The main steps of k-means algorithm are as follows:

(1) Select an initial partition with K=k clusters; repeat steps (2) and (3) until cluster membership stabilizes.

(2) Generate a new partition by assigning each pattern to its closest cluster center.

(3) Compute new cluster centers.

III. FRAMEWORK OF CONTENT-BASED IMAGE RETRIEVAL COMBINED WITH K-MEANS CLUSTERING ALGORITHM

Since image retrieval is according to the similarities between the query image and images in image database, ignoring the similarities between images in image database. The paper applies the clustering algorithm to further explore the similarities between images in image database for reducing the image retrieval space.

In CBIR, Image feature vectors can be represented by a real matrix G, each row gi of which represents the feature vector of an image in database, and each column represents one kind of feature value. Element G(i,k) represents the feature value of the ith image under the kth feature. The relationship among images can be represented by affinity A ¼ (a_{ij}), where aij represents the similarity between the ith image and the jth image. The similarity could be measured by Euclidian distance or other metrics. The image retrieval system combing clustering algorithm is shown as Fig. 1. The system could be any real value symmetric image retrieval system. First, extract the image features of each image in image database and apply the clustering algorithm to analysis the similarities of images in the database for constructing the images clustering database, then, input the query image, extracting its features and comparing the similarities between features of it and those of images in image clustering database, and output the best matching results.

IV. EXPERIMENT RESULT

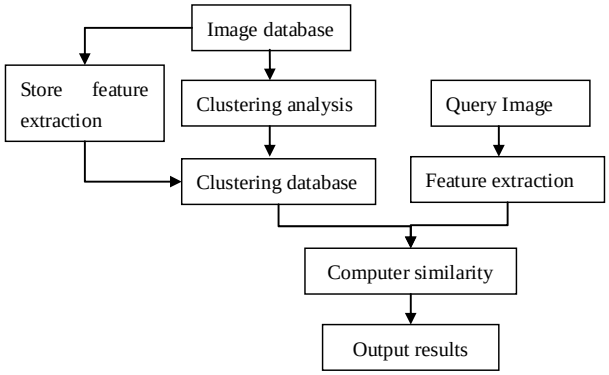


Figure1. Framework of image retrieval system combined with clustering algorithm.

We performed experiments mainly on three images testing set, such as flowers, flags and winter, using color feature only and color and shape feature for retrieval, combining the k-means clustering algorithm with images retrieval.

In order to show k-means clustering algorithm performance in image retrieval system, we design a series of test on the clustering performance. Fig. 2 and Fig.3

show the retrieval results with and without k-means clustering algorithm. The upper left image is the query image. The right part is the retrieved images. Fig. 2 displays the retrieval results without k-means clustering algorithm, while Fig.3 displays the results with k-means clustering algorithm that images similar to each other and to query. In Fig. 2, the query image is a flag and feature vector composed of color and texture. We observe that application k-means clustering algorithm in the images retrieval can throw away some images that are visually irrelevant to the query image for reducing images retrieval space. A leaf image is displayed in the firstly retrieval results in Fig. 2, because of low-level color features. While through doing k-means clustering analysis for image retrieval, such leaf image isn't displayed in the results in Fig.3.

Figure 2. Retrieval results for example query flags image without k-means clustering algorithm

Figure 3. Retrieval results for example query flags image using k-means clustering algorithm before image retrieval.

More query examples are given in Fig. 4 and Fig.5. Fig.4 shows flowers retrieval results without k-means clustering algorithm, and Fig. 5 shows retrieval results using such k-means clustering algorithm to reducing image retrieval space.

V. Conclusion

Image retrieval algorithms always use the similarity between the query image and images in image database. However, they ignore the similarities between images in image database. In this paper we addressed this problem by introducing a graph-theoretic approach for image retrieval post-processing step by finding image similarity clustering to reduce the images retrieving space.

Experiment results show that the efficiency and effectiveness of k-means algorithm in analyzing image

Figure 4. Retrieval results for example query flags image without k-means clustering algorithm

Figure 5. Retrieval results for example query flowers image using k-means clustering algorithm before image retrieval.

clustering, which also can improve the efficiency of image retrieving and evidently promote retrieval precision. This k-means algorithm independent on the feature extraction algorithm is used as a post-processing step in retrieval. The improvements in selecting neighborhood vertices of the retrieval results from tradition image retrieval system in image feature space could also improve the recall rate. Since image features is another problem in image retrieval, finding suitable features is important. Thus, feature selection is a problem for our future work. For k-means clustering algorithm, machine learning and pattern recognition communities need to address a number of issues to improve our understanding of data clustering. Below is about research directions that are worth focusing about such algorithm applications in the image retrieval.

(1) Regardless of the principle (or objective), most clustering methods are eventually cast into combinatorial optimization problems that aim to find the partitioning of data that optimizes the objective. As a result, computational issue becomes critical when the application involves large scale data. For instance, finding the global optimal solution for K-means is NP-hard. Hence, it is important to choose clustering principles that lead to computationally efficient solutions.

(2) A fundamental issue related to clustering is its stability or consistency. A good clustering principle should result in a data partitioning that is stable with respect to perturbations in the data. We need to develop clustering methods that lead to stable solutions.

(3) Choosing clustering principles according to their satisfiability of the stated axioms. Despite Kleinberg's impossibility theorem, several studies have shown that it can be overcome by relaxing some of the axioms. Thus, maybe one way to evaluate a clustering principle is to determine to what degree it satisfies the axioms.

(4) Given the inherent difficulty of clustering, it makes more sense to develop semi-supervised clustering techniques in which the labeled data and (user specified) pair-wise constraints can be used to decide both (i) data representation and (ii) appropriate objective function for data clustering.

ACKNOWLEDGMENT

The work of this paper was supported by the Zhejiang provincial natural science foundation as general science and technology research project (No. Y1080565).

REFERENCES

[1] Shyi-Chyi Cheng , Tian-Luu Wu. Fast indexing method for image retrieval using k nearest neighbors searches by principal axis analysis. S.-C. Cheng, T.-L. Wu / J. Vis. Commun. Image R. 17 (2006) 42–56.

[2] Muhammad Atif Tahir,, Ahmed Bouridane, Fatih Kurugollu. Simultaneous feature selection and feature weighting using Hybrid Tabu Search/K-nearest neighbor classifier. M.A. Tahir et al./Pattern Recognition Letters 28 (2007) 438–446.

[3] Ying Liua, Dengsheng Zhang, Guojun Lu,Wei-Ying Ma. Asurvey of content-based image retrievalwith high-level semantics. Y. Liu et al. / Pattern Recognition 40 (2007) 262 – 282.

[4] Hsin-Chang Yang,Chung-Hong Lee. Image semantics discovery from web pages for semantic-based image retrieval using self-organizing maps. H.-C. Yang, C.-H. Lee / Expert Systems with Applications 34 (2008) 266–279.

[5] Z. Gong, L. Hou U, C.W. Cheang, Web image semantic clustering, in:Proceedings of ODBASE, 2005, pp. 1416–1431.

[6] Y. Dai, D. Cai, Image clustering using semantic tolerance relation model, in:Proceedings on European Internet and Multimedia Systems and Applications, 2007

[7] Z. Hai, Retrieve images by understanding semantic links and clustering image fragments, Journal of Systems and Software 73(2004) 455–466.

[8] L. Duan, Y. Chen, W. Gao, Learning semantic cluster for image retrieval using association rule hyper graph partitioning, in: Proceedings of the 2003 Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing and the Fourth Pacific RimConference on Multimedia, 2003, pp. 1581–1585

[9] C. Zhang, X. Chen, M. Chen, S.-C. Chen, M.-L. Shyu, A multiple instance learning approach for content based image retrieval using one-class support vector machine, in: Proceedings of the IEEE International conference on Multimedia & Expo (ICME), Amsterdam, The Netherlands, 2005, pp. 1142–1145.

[10] F. Jing, C. Wang, Y. Yao, K. Deng, L. Zhang, W.-Y. Ma, IGroup: a web image search engine with semantic clustering of search results, in: Proceedings of ACM MM Demo, 2006.

[11] Ying Liu, Xin Chen, Chengcui Zhang, Alan Sprague. Semantic clustering for region-based image retrieval Journal of Visual Communication and Image Representation, Volume 20, Issue 2, February 2009, Pages 157-166.

[12] Shyi-Chyi Cheng, Tzu-Chuan Chou, Chao-Lung Yang, Hung-Yi Chang.A semantic learning for content-based image retrieval using analytical hierarchy process. Expert Systems with Applications, Volume 28, Issue 3, April 2005, Pages 495-505.

[13] Hsin-Chang Yang, Chung-Hong Lee. Image semantics discovery from web pages for semantic-based image retrieval using self-organizing maps. Expert Systems with Applications, Volume 34, Issue 1, January 2008, Pages 266-279.

[14] Hai Jin, Xiaomin Ning, Weijia Jia, Hao Wu, Guilin Lu.Combining weights with fuzziness for intelligent semantic web search. Knowledge-Based Systems, Volume 21, Issue 7, October 2008, Pages 655-665.

[15] Ying Liu, Dengsheng Zhang, Guojun Lu, Wei-Ying Ma.A survey of content-based image retrieval with high-level semantics. Pattern Recognition, Volume 40, Issue 1, January 2007, Pages 262-282.

[16] Muhammad Atif Tahir, Ahmed Bouridane, Fatih Kurugollu. Imultaneous feature selection and feature weighting using Hybrid Tabu Search/K-nearest neighbor classifier. Pattern Recognition Letters, Volume 28, Issue 4, 1 March 2007, Pages 438-446.

[17] Sarbast Rasheed, Daniel Stashuk, Mohamed Kamel.Adaptive fuzzy k-NN classifier for EMG signals decomposition. Medical Engineering & Physics, Volume 28, Issue 7, September 2006, Pages 694-709.

[18] J. Amores, N. Sebe, P. Radeva.Boosting the distance estimation: Application to the K-Nearest Neighbor Classifier. Pattern Recognition Letters, Volume 27, Issue 3, February 2006, Pages 201-209.

[19] Man Wang, Zheng-Lin Ye, Yue Wang, Shu-Xun Wang. Dominant sets clustering for image retrieval. M. Wang et al. / Signal Processing 88 (2008) 2843–2849.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 278-280

Design and Implement of Distributed Document Clustering Based on MapReduce

Jian Wan, Wenming Yu[1], and Xianghua Xu
Grid and Services Computing Lab
School of Computer Science and Technology
Hangzhou Dianzi University, Hangzhou 310037, China
[1]yuwenming_001@163.com

***Abstract*—In this paper, we describe how document clustering for large collection can be efficiently implemented with MapReduce. Hadoop implementation provides a convenient and flexible framework for distributed computing on a cluster of commodity machines. The design and implementation of tfidf and K-Means algorithm on MapReduce is presented. More importantly, we improved the efficiency and effectiveness of the algorithm. Finally, we give the results and some related discussion.**

***Index terms*—MapReduce, tfidf, K-Means clustering**

I. INTRODUCTION

With the rapid development of the Internet, huge volumes of documents need to be processed in a short time. Research on web mining focuses on scalable method applicable to mass documents[1]. Storage and computing of mass documents data in a distributed system is an alternative method[2]. In distributed computing, a problem is divided into many tasks, each of which is solved by one computer. However, many problems such as task scheduling, fault tolerance and inter-machine communication are very tricky for programmers with little experience with parallel and distributed system.

In this paper we describe our experiences and findings of document clustering based on MapReduce. MapReduce [3] is a framework which programmers only need to specify Map and Reduce functions to make a huge task parallelize and execute on a large cluster of commodity machines. In the document pre-processing stage, we design a new iterative algorithm to calculate tfidf weight on MapReduce in order to evaluate how important a term is to a document in a corpus. Then, a K-Means clustering is implemented on MapReduce to partition all documents into k clusters in which each documents belongs to the cluster with the same meaning. More importantly, we find that ignoring the terms with the highest document frequencies can not only speed up our algorithm on MapReduce, but also improve the precision of document clustering slightly. Experiments show that our method in approximately linear growth in required running time with increasing corpus size for corpus containing several ten thousand documents.

II. MAPREDUCE AND HADOOP

Many real world tasks have the same characteristics: a computation is applied over a large number of records to generate partial results, which are then aggregated in some fashion. MapReduce is a programming model which is specializing in handing problems having "Divide and Conquer" structure. MapReduce inspired by functional language consists of the Map and Reduce abstract concepts. A map function process each logical "record" in our input in order to compute a set of intermediate key/value pairs, and then a Reduce function accepts an intermediate key and a set of values for that key in order to combine the derived data appropriately. Figure 1 illustrates the two processing stages.

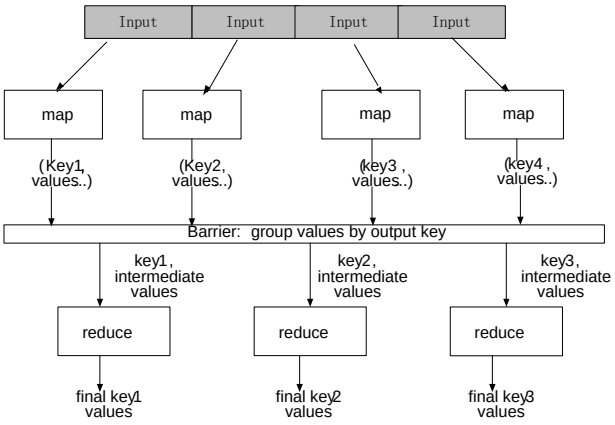


Figure1. Processing procedure of MapReduce

Apache Hadoop [4] is a Java software framework that consists of a MapReduce model and a distributed file system (HDFS, similar to GFS[5]). HDFS is designed to scale to mass storage across multiple machines, and transparently provides read/write, backup and fault-tolerance for users. Hadoop is becoming increasingly popular[6, 7] because it hides the messy details of parallelization, fault-tolerance, data distribution and load balancing.

III. CALCULATE TFIDF ON MAPREDUCE

The tfidf weight [8](term frequency–inverse document frequency) is a weight often used in text mining and information retrieval. The importance of a term increases proportionally to the number of times a term appears in the document but is offset by the frequency of the term in the corpus. Therefore the weight of feature term t_i in the corpus can be calculated using the classic tfidf scheme in formula (1).

$$w_{ij} = tf_{ij} \times idf_i = |t_i| / |d_j| \times \log(N / n_i) \qquad (1)$$

f_{ij} is the frequency of feature term t_i in the document d_j. It can be designated $|t_i|/|d_j|$, where $|t_i|$ is the number of occurrences of the term t_i in document d_j, and $|d_j|$ is the total number of terms in document d_j. N is the total number of documents in the corpus, and n_i is the number of documents that contains term t_i. We can see from the formula (1) that we should calculate $|t_i|$, $|d_j|$ and n_i on MapReduce to get tfidf.

1) Number of times a term t_i appears in a given document: The format of input data to Map function is (Docname, content), which means that document name is key and relevant content is value. For each term in the document, Map function output ((term, Docname), 1) which means this term occurrences one time in this document. Reduce functions accepts the output of former Map functions, and aggregate the records with the same key. The output format of Reduce functions is ((term, Docname), $|t_i|$). In practice, we can add a Combiner function to accelerate the computing speed. The function of Combiner function is the same as the Reduce function.

2) Number of terms in each document: This step's input data is the output of the first step, and the map functions convert the format into (Docname, (term, $|t_i|$)). The Reduce functions get the records sharing the same docname, and accumulate the number of different terms $|t_i|$ into $|d_j|$ in the same document. The output format of this step is ((term, Docname), ($|t_i|$, $|d_j|$))。

3) Number of documents term t_i appears in: The Map function in this step turn the output of above step into the format of (term, (Docname, $|t_i|$, $|d_j|$, 1)), which means that this term appears in one document. The Reduce function accumulate "1" with the same term into n_i, this is the number of documents contain the term t_i。 The output format of this step is ((term, Docname), ($|t_i|$, $|d_j|$, n_i)).

4) *Calculate tfidf :* The output of step 3 means that $|t_i|$ is the occurrence time of term t_i in document d_j, $|d_j|$ is the number of all terms in document d_j and n_i is the number of documents contain the term t_i. We can just use formula (1) to calculate tfidf of terms in different documents. The output format of the result is $(Docname, (term_1 \& tfidf_1, \cdots, term_n \& tfidf_n))$.

IV. K-MEANS CLUSTERING

K-Means clustering [9] choose k initial points and mark each as a center point for one of the k sets. Then for every item in the total data set it marks which of the k sets it is closest to. It then finds the average center of each set, by averaging the points which are closest to the set. With the new set of centers (centroid), it repeats the algorithm until convergence has been reached.

The implementation of document clustering on MapReduce accepts two input directories: one is the documents directory with the output of calculating tfidf, and one is centers directory with k initial document centers. The k initial document centers are chosen from the records of documents directory. Note that the k-line document data have the same terms as fewer as possible.

In every iteration, the MapReduce framework will partition the input files of document directory into a set of M splits, and then these splits are processed in parallel by M Map functions. Map functions read in a document with the format $(Docname, (term_1 \& tfidf_1, \cdots, term_n \& tfidf_n))$. Map functions should determine which of the current set of k document centers (in centers directory) the document is closest to and emits a record containing all the document's data and its chosen k-center with the format $(k-center, (Docname, term_1 \& tfidf_1, \cdots, term_n \& tfidf_n))$. The Reduce function receives a k-center and all documents which are bound to this k-center. It should calculate a new k-center, and put the new k-center in centers directory. To evaluate the distance between any two documents, we use the cosine similarity metric of tfidf, and use arithmetic average to calculate the new k-center. Note that the contents in document directory will not change during the process. The whole K-Means clustering on MapReduce can be expressed as the following two step:

$$map: (Docname, term \& tfidf\_list) \rightarrow (k\_center, (Docname, term \& tfidf\_list)) \quad (2)$$

$$reduce: (k\_center, (Docname, term \& tfidf\_list)) \rightarrow (new\_k\_center, term \& tfidf\_list) \quad (3)$$

V. EXPERIMENT EVALUATIONS

In our experiment, we used Hadoop version 0.18.2 running on a cluster with 5 machines (one is the master, also act as a slave). Each machine has two single-core processors (running at 2.33GHz), 1GB memory, and 130GB disk, and the network bandwidth is 100Mbps. The configuration of Hadoop is two map functions running on a processor core simultaneously. The number of map functions can be controlled precisely.

We used the Sogou documents classification corpus[1], containing 80k documents, totaling approximately 211MB. There are 10 different subjects in the corpus, and 8k documents in each subject. These documents are parsed and terms are stemmed. All empty words (we maintain a Chinese stop word list) in these documents are removed.

We always use running time to measure the efficiency. One issue that became evident in initial experiments was the prevalence of "stragglers", which means one or two reducers that take significantly longer

[1] http://www.sogou.com/labs/dl/c.html

than the others (this is a common problem, see) due to the Zipfian distribution of terms. In our experiment, we eliminate the terms with the highest document

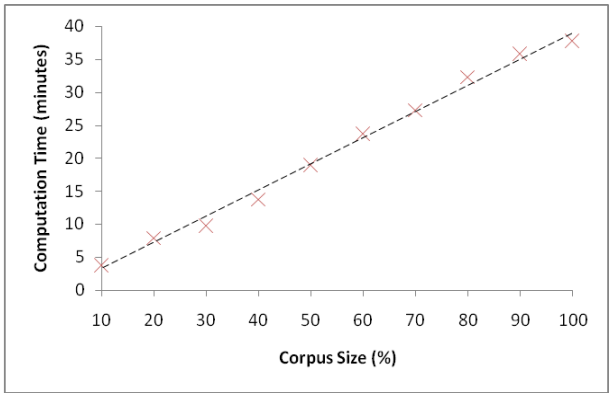


Figure2. Running time of tfidf algorithm with the different size of collections

frequencies. We adopt a 95% cut at step 3 in section 3, which means that the most frequent 5% of terms were ignored. This method greatly increases the efficiency of our algorithm on Hadoop. On the other hand, because the terms we discarded are non-discriminative, the precision of document clustering is improved slightly.

Figure 2 shows the running time of tfidf algorithm on our cluster with increasing collection size for collection containing 80k documents. We get the result just to see the effect of our algorithm on Hadoop, don't configure our cluster in optimizing. We find the time used for calculating the whole collection is more than half an hour. However, the running time and space required is approximately linear with the size of collection, this is characteristics we expect in processing mass data.

We measure the running time of K-Means clustering on the clusters, and then implement conventional K-Means on single machine as a benchmark. These results were compared against an equivalent run on the machine with the same times of iterations (5 times). We can see from Figure 3 that the running time of K-Means algorithm on a cluster with 5

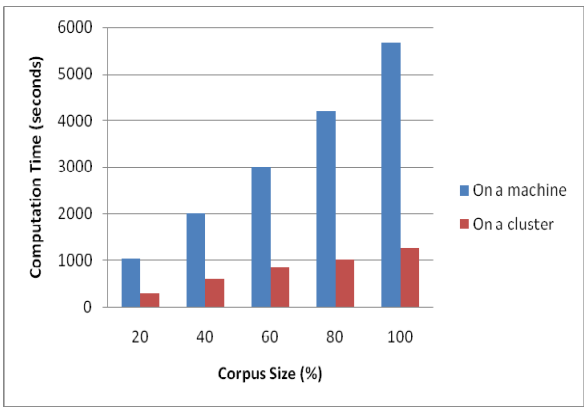


Figure 3. Running time comparison of K-Means on different size of collections

machines relatively much less than on a single machine. More importantly, with the size increasing of our collections, the advantage of efficiency has become increasingly evident.

VI. Conclusions

The paper has introduced a mass documents clustering in a distributed system Hadoop. Experiments show the scalability of our method in processing mass data. The contributions of our work lie in both design and implement tfidf and K-Means algorithm on MapReduce. We believe that our work provides an example of a programming paradigm that could be useful for a broad range of text analysis problems. Finally, we always pay attention to the alternative approaches to similar problems based on MapReduce [10]. Hadoop provides unprecedented opportunities for researchers to handle real-world problems at scale.

Acknowledgment

This paper is supported by National Science Foundation of China under grant No.60873023, and Science and Technology R&D Program of Zhejiang Province, China under grant No. 2008C13080, No.2007C21G3230005.

References

[1] Roberto, J.B., M. Yiming, and S. Ramakrishnan, Scaling up all pairs similarity search, in Proceedings of the 16th international conference on World Wide Web. 2007, ACM: Banff, Alberta, Canada.

[2] Michael, J.F., Evolution of distributed computing theory: from concurrency to networks and beyond, in Proceedings of the twenty-seventh ACM symposium on Principles of distributed computing. 2008, ACM: Toronto, Canada.

[3] Jeffrey, D. and G. Sanjay, MapReduce: simplified data processing on large clusters. Commun. ACM, 2008. 51(1): p. 107-113.

[4] Apache Lucene Hadoop[EB/OL].http://hadoop.apache.org/.

[5] Sanjay, G., G. Howard, and L. Shun-Tak, The Google file system, in Proceedings of the nineteenth ACM symposium on Operating systems principles. 2003, ACM: Bolton Landing, NY, USA.

[6] Jimmy, L., Brute force and indexed approaches to pairwise document similarity comparisons with MapReduce, in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. 2009, ACM: Boston, MA, USA.

[7] ZHENG Xin-jie, ZHU Cheng-rong, and X. Qi-bang, Design and Implementation of Distributed Ray Tracing Using MapReduce. Computer Engineering, 2007(22).

[8] Salton, G. and T.Y. Clement, On the construction of effective vocabularies for information retrieval, in Proceedings of the 1973 meeting on Programming languages and information retrieval. 1973, ACM: Gaithersburg, Maryland.

[9] Fasulo, D., An Analysis of Recent Work on Clustering Algorithms, in Technical Report UW-CSE-01-03-02. 1999, ACM: University of Washington.

[10] Chu, C.-T., S.K. Kim, and Y.-A. Lin, Mapreduce for machine learning on multicore, in In Proceedings of Neural Information Processing Systems Conference (NIPS) 2007.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 281-284

Analysis and Application of Iteration Skeletonization Algorithm in Recognizing Chinese Characters Image

Tingmei Wang[1], Ge Chen[2], and Zhansheng Chen[2]
[1] Applied Science and Technology Institute Beijing Union University, Beijing, China
Email: wtm9329@yahoo.com.cn
[2] Applied Science and Technology Institute Beijing Union University, Beijing, China
Email: xxt_chenge@buu.edu.cn, zenithcoup@sina.com

***Abstract*—The paper studied several image skeleton extraction algorithms, such as Zhang-Suen, Rosenfeld and Pavlidis etc. And compared extraction effects based on different Chinese characters by making computer programs, and gave advice how to choose the best algorithm for recognizing a certain Chinese characters image, and then gave a conclusion that Zhang-Suen and Rosenfeld algorithms were the best algorithms for extracting Chinese characters skeletons.**

***Index Terms*—Chinese Characters Image Processing; Image Skeleton Extraction; Zhang-Suen; RosenFeld; Pavlidis**

I. INTRODUCTION

With development of Internet and application of computer technology, a plenty of information exists in form of images. Retrieving among so many images becomes very difficult. When retrieving Image based on key words, we must add mark for them so that efficiency of retrieval is very low. In this case, image retrieval based on content (CBIR) which use color, strip, shape, space relation etc, is needed. Image retrieval based on space relation uses space topology structure of objects in images to retrieve. Image skeleton is one of the best effective ways to represent the topology structure. It is widely used in shape description, pattern recognition, and industrial inspection, and image compression coding etc. Image skeleton was proposed by Blum [1]. He represented image skeleton by means of axis.

Image skeleton extraction algorithm generally possesses the following characteristics: The first is connectivity, which means connectivity of image skeleton must be consistent with original image. The second is thinning, which means width of image skeleton should be one pixel. The third is axis, which means image skeleton should be as far as possible the center line of original image. The fourth is maintainability, which means image skeleton should retain as far as possible detail of original image. Last one is rapidity, which means running speed of algorithm should be as fast as possible.

Image skeleton extraction in Chinese character recognition is an important research topic. Image skeleton extraction algorithm consists of iterative algorithm and non-iterative algorithm. This paper mainly discusses several classical iterative skeleton extraction algorithms in the character image skeleton extraction applications.

II. MODELS OF IMAGE SKELETON EXTRACTION

A. Fire Spreading

Fire spreading refers to all border points of image are lighted at the moment of t=0, the flame spreads towards internal of image at the same speed. When the meet happens in front of wave, the flame goes out. Collection of points where flame goes out constitutes axis named image skeleton denoted by SKF shown as fig.1

B. Maximum Disc

Maximum disc indicates that suppose D is one of inscribed disc of image, this means at least two points of the disc are tangent to image edge. If D is not a subset of other inscribed discs within image, D is the greatest disc as shown in fig.2. Now image skeleton denoted by SKM can be defined as collection of the greatest discs within image.

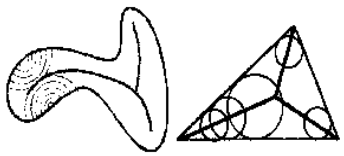


Figure 1. Fire spread skeleton. Figure 2. Maximum disc skeleton

C. Peeling Method

Peeling method indicates that staring from edge of line, you peels it off one layer with one pixel width at a time till the single pixel lines that are connected each other and constructed by single raster are obtained. One condition should be considered during peeling because a line in different locations may have different widths, that is to say the pixel that may result in lines disconnected can never be peeled. The basic principle of the algorithm is that peeling pixels that have no effect on connectivity of image raster topology, and conversely these pixels should be retained.

III. ANALYSIS OF SEVERAL CLASSICAL IMAGE SKELETON EXTRACTION ALGORITHMS BASED ON PEELING

AP-PROC-CS-09CN005

A. *Principle of Template Matching Algorithm*

Eight points adjacent to pixel p are called 8-neighborhood of pixel P as shown in fig.3, and it can be denoted by Pi such that the value of i is less than 1 and greater than 8. P1,P3,P5 and P7 are called four-adjacent area point, and P1,P2,P3,P4,P5,P6,P7,P8 are called eight-adjacent area point.

Template matching algorithm uses a template 3*3 to retrieve images and judges if there are points being the same case as the template. If there are such points, you should execute corresponding operations according to the operations of template such as retaining operation, deletion etc. In short, Rosenfeld, Zhang-Suan and Pavlidis algorithms are image skeleton extraction algorithm based on template matching.

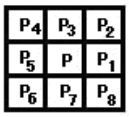


Figure 3. 8- neighborhood of pixel P.

B. *Rosenfeld Algorithm*

Rosenfeld algorithm was first advanced by Stefanelli R and Rosenfeld A in 1971[2].It is an image skeleton extraction algorithm based on parallel template matching. The algorithm defines condition for final pixel as shown in fig.4 and judges if edge pixel satisfies the condition so as to determine how to operate it. As shown in fig.4, there exists at least one black point among these points marked X, and the same is true for Y. here, black point represents foreground color, and white point represents background color, and gray point represents foreground or background colors. You should retain pixel points satisfying final pixel condition, otherwise remove them. Processing cycle of the algorithm consists of four main steps: completing a removal of edge pixels from four directions-moving up, down, left or right, then repeating this process until no deleting occurs in a certain cycle. Now process of image skeleton extraction is finished.

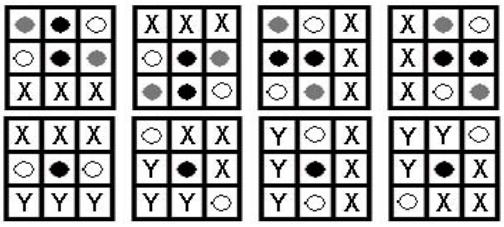


Figure 4. Final pixel condition

Image skeleton extraction procedure can be described as follows:

```
while(there still exist pixels to be deleted)
{
    For foreground color point P search:
    {
      If 8-neighborhood of the upper edge point of P matches templates 2,3,4,5,6,7,8, continue scanning.
      If 8-neighborhood of the lower edge point of P matches templates 1,3,4,5,6,7,8, continue scanning.
      If 8-neighborhood of the left edge point of P matches templates 1,2,4,5,6,7,8, continue scanning.
      If 8-neighborhood of the right edge point of P matches templates 1,2,3,5,6,7,8, continue scanning.
      Otherwise mark point P for deleting.
    }
}
```

C. *Zhang-Suen Algorithm*

Zhang-Suen algorithm was advanced first by Zhang T Y in 1984 [3]. It is also an image skeleton extraction algorithm based on parallel template matching and deletion. The algorithm defines condition for points to be deleted as shown in fig.5. For those points satisfying deleting condition, you should delete them. As shown in fig.5, if satisfying one black point or two gray points representing background color, you delete the point. Processing cycle of the algorithm consists of two steps. Finding a point matching with the template shown in fig.5(1) is the first step, the second step is to verify the point matching with the template shown in fig.5(2),then repeating the above process until there are no points to be

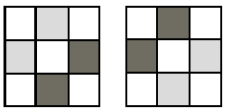


Figure 5. Zhang-Suen chinning template

deleted in a certain cycle.

Image skeleton extraction procedure of Zhang-Suen algorithm can be described as follows:

```
while (there still exist points to be deleted)
{
    For each foreground point P search:
    {
      If 8-neighborhood of edge point of P matches template 1, delete point P.
      If eight adjacent area of edge point of P matches template 2, delete point P.
    }
}
```

D. *Pavlisis Algorithm*

Palidis algorithm was advanced first by Pavlidis in 1982[4]. It is an image skeleton extraction algorithm based on template matching and retaining. The algorithm defines the condition for points to be retained showed in fig.6. For those points satisfying retaining condition, you should retain them. Processing cycle of the algorithm consists of four steps. The first step is to judge if point of right edge matches the template, then repeat the same operation to process its upper edge, left edge and lower edge until there are no points to be deleted. Now process of image skeleton extraction is finished.

Image skeleton extraction of Pavlidis was described as follows:

```
while(there still exist points to be deleted)
{
```

For each foreground point search:

{

If 8-neighborhood of right edge point of P matches templates 1,2,3,4,5,6, continue scanning.

If 8-neighborhood of the upper edge point of P matches templates 1,2,3,4,5,6, continue scanning.

If 8-neighborhood of the left edge point of P matches templates 1,2,3,4,5,6, continue scanning.

If 8-neighborhood of lower edge point of P matches templates 1,2,3,4,5,6, continue scanning.

Otherwise mark the point P for deletion.

}

}

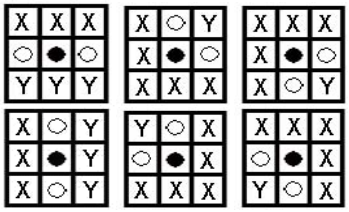


Figure 6. The condition of retaining point(At lease one of X,Y is 1)

E. Other Image Skeleton Extraction Algorithms

The deleting condition of Hilditch algorithm and retaining condition of Pavlidis algorithm are complementary [5]. Similar image skeleton using these two algorithms can be obtained, so their description will not be discussed in detail. Naccache algorithm marks safety point including either breakpoint or endpoint to be retained and otherwise these points are to be removed, then deletes all the points described as removed until there is no point to be deleted.

Deutsch algorithm is an image skeleton extraction algorithm based on deletion. It defines some deletion conditions, and if there exist such points that satisfy deletion condition, the points should be deleted. Otherwise the points should be retained. The algorithm extracts image skeleton by two cycles executed alternately.

F. Comparison of Skeleton Extraction Algorithms

From the discussion presented above, we can know that these algorithms mentioned obtain image skeleton by conditions matching and many times iterations. Some of them extract image skeleton through given conditions, such as Hilditch algorithm and Deutsch algorithm. Some of them extract image skeleton through templates matching, such as Zhang-Suen algorithm and Rosenfeld algorithm. Hilditch algorithm, Deutsch algorithm and Zhang-Suen algorithm extract image skeleton based on deletion, and Rosenfeld algorithm, Pavlidis algorithm and Naccache algorithm extract image skeleton based on retaining. Though reservation conditions are called by different names in these algorithms, final pixel condition in Rosenfeld, retaining point condition in Pavlidis, and safety point condition in Naccache, but their principles are the same.

All the skeleton extraction algorithms mentioned above are based on peeling. In other words, we can obtain image skeleton of single pixel connecting by peeling edge pixels many times. These algorithms have different times of iteration. One iteration cycle of Pavlidis algorithm and Rosenfeld algorithm consists of four steps, and Zhang-Suen algorithm consists of two steps in one iteration cycle. At the same time, different iteration processes lead to different number of templates. Zhang-Suen algorithm has the least number of templates so that its speed is the fastest. The number of templates and the times of iterations have effect on efficiency of algorithms. The less the algorithm uses the number of templates and the less the algorithm has the times of iteration, the faster the algorithm runs.

IV. EXPERIMENT RESULT AND ANALYSIS

Programming and running these algorithms including Hilditch, Pavlidis, Rosenfeld, Naccache, Deutsch and Zhang-Suen in turns to extract image skeleton for handwritten Chinese characters, printed Chinese characters and Chinese calligraphy as shown in fig.7 and processing results are shown in fig.8,fig.9 and fig.10.

Figure 7. Experiment images

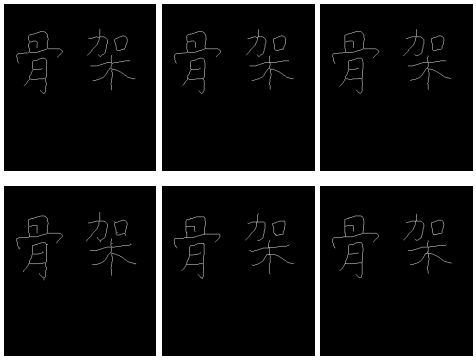


Figure 8. Handwritten Chinese characters

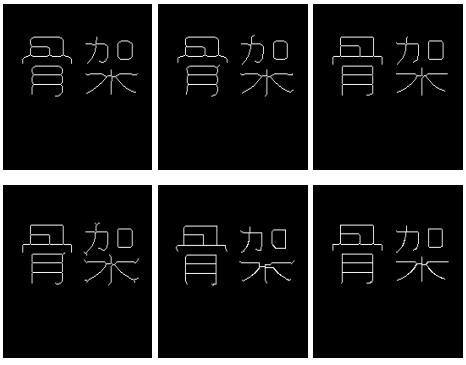


Figure 9. Printed Chinese characters

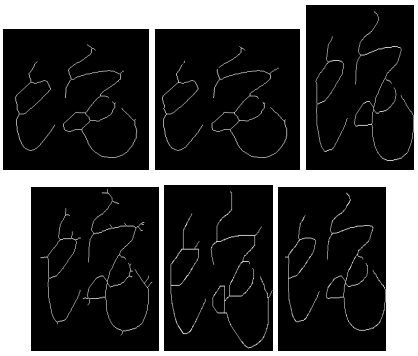


Figure 10. Chinese calligraphy

Drawing a comparison for experimental figures, conclusion is shown as follows:

- If you want to recognize Chinese calligraphy, handwritten Chinese characters and printed Chinese characters, the processing effect of Rosenfeld algorithm is better than Hilditch, Pavlidis, Daccache and Deutsch algorithms'.
- Both Hilditch algorithm and Pavlidis algorithm can produce similar image skeletons. They can result in good effect for processing handwritten Chinese characters and printed Chinese characters. But when it comes to processing Chinese calligraphy, they can not process the intersection of lines accurately.
- Considering image skeleton connectivity, Pavlidis algorithm is worse than Hilditch. When processing Chinese calligraphy by Pavlidis algorithm, image skeleton exists apparent fracture shown as fig.11.
- Using Rosenfeld and Zhang-Suen algorithms, we can obtain image skeleton with no superfluous branches, and they can process intersection of lines accurately.
- When processing handwritten Chinese characters and printed English characters, the processing effect of Naccache algorithm is better than that of others'. But for processing printed Chinese characters and Chinese calligraphy, the effect is not satisfactory, because there exist so many superfluous branches.
- When processing handwritten Chinese characters, Deutsch algorithm can get good effect, but if processing printed Chinese characters and Chinese calligraphy, the effect of it is not satisfactory, because there are not only many superfluous branches, but also a little deformity.

V. CONCLUSIONS

The paper introduces some classical models of image skeleton extraction, analyzes principles of several classical image skeleton extraction algorithms based on peeling and tests these algorithms through processing Chinese calligraphy, handwritten Chinese characters and printed Chinese characters. Experiment result shows that efficiency of Zhang-Suen algorithm is higher than that of other algorithms. Using Rosenfeld and Zhang-Suen algorithms to process Chinese characters, we can obtain image skeleton with a good topology features and without superfluous branches. So a conclusion can be drawn that Zhang-Suen and Rosenfeld algorithms are the best algorithms for extracting Chinese characters skeletons.

REFERENCES

[1] Blum H. A transformation for extracting new description of shape. In: Wathen-Dunn W ed. Model for the Perception of Speech and Visual. Cambride, Massachusetts: MIT Press, 1967,pp 362-380.

[2] Stefanelli R, Rosenfeld A.Some parallel thinning algorithms for digital pictures[J].Journal of ACM,1971(2).

[3] Zhang T Y. A fast parallel algorithm for thinning digital patterns[J]. Communications of ACM, 1984, 27(3):236-239.

[4] Pavlidis T.Algorithms for Graphics and lmage Procesing[M] . Rockville:Computer Science Press,1982.

[5] Hilditch C J.Linear Skeletons from Square Cupboards[J].Machine Intelligence, 1969;4:403.

[6] Naccache N J, Shinghal R. SPTA: A proposed algorithm for thinning by binary patterns[J]. IEEE Transactions on System, Man Cybernetics, 1984, 14(3):409-418.

[7] Deutsch E S. Thinning algorithms on rectangular hexagonal and triangular arrays[J].Communications of the ACM,1972,15(9):827～837.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 285-288

A Multi-attribute Assessment Method for E-Commerce Risks

Caiying Zhou [1], and Longjun Huang [2]
[1] Jiangxi University of Science and Technology,Ganzhou, China
zhoucaiying_007@ 163.com
[2] JiangXi Normal University ,NanChang,China
Huanglong@ jxnu.edu.cn

***Abstract*—Based on multi-attribute decision making theory, a multi-attribute assessment framework for E- commerce risks is presented. The paper applies an improved evidential reason- ing(ER) approach to support evaluation decision, and uses a credit interval to figure scoring result and allow something absence. E-Commerce risks includes environmental risks, management risks, technical risks and information risks. A numerical example of E-Commerce risks evaluation problem is discussed to demonstrate the implementation of multi-attribute assess-ment approach.**

***Index Terms*—multi-attribute assessment method, E-commerce risks, evidential reasoning approach**

I. INTRODUCTION

With universal access to the internet, internet-based E-Commerce (EC) came into being. Compared with traditional commerce, E-Commerce has the characteristics of high efficiency, convenience, integrated and scalability. But through browser / server mode, E-Commerce which achieve the consumer's online shopping, online transactions between merchants and online e-payments is based on an open network environment, its security compared to traditional methods in terms of traditional business approach is particularly prominent.

Quite a few scholars studied the E-Commerce risks. Greenstein M. defined E-Commerce risk as follows: the possibility of confidential data loss or of hardware damage or result in physical or financial harm on the other side of the destruction, establishment or use of data or programs[1].Greenstein M. defined the risk of E-Commerce from the point of view of data storage and transmission which belongs to the technical risk areas in this paper later mentioned. In the domestic scholars, Luan Xin-xin explored the management of E-Commerce technology [2]. Liu Wei-jiang, Wang Yong studied E-Commerce risks and control strategy [3].Zhu Ya-shu, Zhu Jin-zhou researched network payment risk and prevention [4].Zhang Xin analyzed the development of firewall technology [5].In fact, E-Commerce risks mainly refers to self-inflicted losses of people who use various terminal devices through a variety of networks (such as the Internet, wireless Internet, etc.) in the course of their transaction.

The starting point for e-commerce security practices are risks evaluation, when a potential threat attacks system's vulnerability points, there will produce risks and lead to the system destruction and damage. Risks evaluation is the process of interpretation and analysis of the risks. Chinese scholars have done related study of the risks evaluation of E-Commerce. For instance, Zhao Dong-mei researched the fuzzy assessment of network security risks based on entropy weight coefficient method [6].Lin Bin and Du Hong-mei developed the risks assessment based on the probability distribution method [7]. E-Commerce risks are the basis of the realization of e-commerce. Analysis of E-Commerce risks is an important research topic to protect the security of e-commerce. In this regard, this paper presents a model of E-Commerce risks based on the multi-attribute evaluation method to make a quantitative analysis of the E-Commerce risks.

II. MODEL CONSTRUCTION

A. Evaluation Framework

In the basis of using the available study results in E-Commerce risks, according to objectivity and functional principle, a new multi-attribute assessment framework of E-Commerce risks is proposed [8-11], which includes four general attributes: environmental risks, management risks, technical risks and information risks. Every general attributes can be divided into several basic attributes, see Table I. ω_i and ω_{ij} in the parentheses are the relative weights of the general attributes and basic attributes with $0 \le \omega_i$, $\omega_{ij} \le 1$ which may be estimated by using existing methods such as simple rating methods or more elaborate methods based on the pair wise comparisons of attributes[12],and the paper does not research them.

B. Scoring Method

According to the complexity of the evaluation attributes, a reasonable method is to constitute an Experts Group, and the experts grade the basic attributes after the

investigation about the companies. In E-Commerce risks evaluation, we need to deal with numerical data and qualitative information with uncertainty. Now evaluation grades set is defined as follows:

$$H = \{H_1, H_2, H_3, H_4, H_5\} \quad (1)$$

={ Poor,Indifferent,Average,Good,Excellent }

A given evaluation for e_i of an alternative may be mathematically represented as the following distribution:

$$S(e_i) = \{(H_n, \beta_{n,j}) \mid n = 1,\ldots,5\}, i = 1,\ldots, L \quad (2)$$

Where $\beta_{n,i}$ denotes a degree of belief and $\beta_{n,i} \geq 0$, $\sum_{n=1}^{N} \beta_{n,i} \leq 1$ .The above distributed assessment reads that the attribute e_i is assessed to the grade H_n with the degree of belief $\beta_{n,i}$, n= 1 ,… N. An assessment $S(e_i)$ is complete if $\sum_{n=1}^{N} \beta_{n,i} = 1$ and incomplete if $\sum_{n=1}^{N} \beta_{n,i} < 1$.

TABLE I. E-COMMERCE RISKS EVALUATION FRAMEWORK

First Level	Second Level	Third Level	
E-Commerce Risks	ER[a] (ω_1 =0.25)	ER1(ω_{11} =0.2)	Financial Risk
		ER2 (ω_{12} =0.3)	Industrial Risk
		ER3 (ω_{13} =0.2)	Legal Risk
		ER4 (ω_{14} =0.3)	Credit Risk
	MR[b] (ω_2 = 0.25)	MR1(ω_{21} =0.4)	Steady Management System
		MR2 (ω_{22} =0.6)	Implementation of Management Systems
	TR[c] (ω_3 = 0.3)	TR1(ω_{31} =0.4)	Infrastructure Risk of Information Technology
		TR2 (ω_{32} =0.3)	Data Access Technical Risk
		TR3 (ω_{33} =0.3)	Online Payment Technology Rrisk
	IR[d] (ω_4 = 0.2)	IR1 (ω_{41} =0.25)	Asymmetric Information Risk
		IR2 (ω_{42} =0.25)	Information Imperfect Risk
		IR3 (ω_{43} =0.25)	Risk of Misappropriation and Abuse of Information
		IR4(ω_{43} =0.25)	Information Lag Risk

a. ER represents Environmental Risk; b. MR represents Management Risk

c. TR represents Technical Risk; d. IR represents Information Risk;

As for qualitative attributes, assessors' subjective judgment has a great influence on assessment results. Assessors' experience, individual preference, understanding of the evaluation criteria all impact the assessment result's accuracy and consistence. Suppose there are s experts who grade attribute e_i , and $s_{n,j}$ ($s_{n,j} \leq s$) experts consider e_i belonging to H_n evaluation grade, then $\beta_{n,i} = S_{n,j} / S$ .There are some experts don't grade e_i if $\sum_{n=1}^{S} S_{n,j} / S$ .For example, an assessment result of some attribute e_i is $\{H_3(0.5), H_4(0.4)\}$, which denotes that 50% sure that the attribute is average and 40% sure that it is good and the missing 10% represents the degree of ignorance or uncertainty.

As for quantitative attributes, the values of the attributes are numerical and objective. The experts don't need to determine the data of the attributes, only need to determine the equivalence rule. Interval $[h_{n,i}^{-}, h_{n,i}^{+}]$ is equivalent to H_n , let $f(e_i)$ denote the value of the attribute e_i ,there are two possibility:

if $h_{n,i}^{-} \leq f(e_i) \leq h_{n,i}^{+}$,then $\beta_{n,i} = 1$,

$\beta_{k,i} = 0 (k \neq n)$; And if $h_{n-1,i}^{+} \leq f(e_i) \leq h_{n,i}^{-}$, then

$$\beta_{n-1,i} = (f(e_i) - h_{n-1,i}^{+}) / (h_{n,i}^{-} - h_{n-1,i}^{+})$$

$$\beta_{n,i} = (h_{n,i}^{-} - f(e_i)) / (h_{n,i}^{-} - h_{n-1,i}^{+})$$

$$\beta_{n,i} = 0 (k \neq n-1, n)$$

So qualitative attributes and quantitative attributes can be described in the form of formula (1).The scoring method uses interval to instead of numerical score expresses the assessment distribution, and make incomplete evaluation with missing evidence and information uncertainty.

III. DATA ANALYSIS METHOD

A. ER Approach in MADA

Based on the D-S (Dempster-Shafer) theory [13], Yang and Singh studied that the evidential reasoning(ER) approach was used in multiple attribute decision analysis (MADA)[14], Yang, Van Nam Huynh improved the ER approach in later researches, and the paper used the improved ER approach.

Suppose there is a simple two-level hierarchy of attributes with a general attributes at the top level and a number of basic attributes at the bottom level. Suppose there are L basic attributes e_i (i= 1… L) associated with a general attribute y . Define a set of L basic attributes as follows:

$E = \{e_1, e_2, ..., e_L\}$ .Suppose the weights of the attributes are given by ω_i ,which denotes the relative weight of the i th basic attribute e_i with $0 \leq \omega_i \leq 1$.

Let $m_{n,i}$ be a basic probability mass representing the degree to which the i th basic attribute e_i supports the

hypothesis that the attribute y is assessed to the n th grade H_n . Let $m_{H,i}$ be a remaining probability mass unassigned to any individual grade after all the N grades have been considered for assessing the general attribute as far as e_i is concerned.

$m_{n,i}$ and $m_{H,i}$ are calculated as follows:

$$m_{n,i} = \omega_i \beta_{n,i} \ , \ n = 1,\ldots, N \quad (3)$$

$$m_{H,i} = 1 - \sum_{n=1}^{N} m_{n,i} = 1 - \omega_i \sum_{n=1}^{N} \beta_{n,i}, \quad (4)$$

Define $E_{I(i)} = \{e_1, e_2, ..., e_i\}$ as the subset of the first basic attributes. Let $m_{n,I(i)}$ be a probability mass defined as the degree to which all the i attributes in $E_{I(i)}$ support the hypothesis that y is assessed to the grade H_n . $m_{H,I(i)}$ is the remaining probability mass unassigned to individual grades after all the basic attributes in $E_{I(i)}$ have been assessed. $m_{n,I(i)}$ and $m_{H,I(i)}$ can be generated by combining the basic probability masses $m_{n,j}$ and $m_{H,j}$ for all $n = 1,\ldots, N$, $j = 1,\ldots i.$

$$\text{Let}, 0 \le \omega \le 1, \sum_{i=1}^{L} \omega_i = 1 \ (i = 1,\ldots L.),$$

$$m_{H,i} = \bar{m}_{H,i} + \tilde{m}_{H,i}, \bar{m}_{H,i} = 1 - \omega_i,$$

$$\tilde{m}_{H,i} = \omega_i (1 - \sum_{n-1}^{N} \beta_{n,i})$$

The following ER algorithm is the first i assessments with the $(i+1)$th assessment using the same process in a recursive manner:

$$m_{n,I(i+1)} = K_{I(i+1)} (m_{n,I(i)} m_{n,i+1} + m_{n,I(i)} m_{H,i+1} + m_{H,I(i)} m_{n,i+1}) \ , n = 1,\ldots, N \quad (5)$$

$$\tilde{m}_{H,I(i+1)} = K_{I(i+1)} (\tilde{m}_{H,I(i)} \tilde{m}_{H,i+1} + \bar{m}_{H,I(i)} \tilde{m}_{H,i+1} + \tilde{m}_{H,I(i)} \bar{m}_{H,i+1}) \quad (6)$$

$$\bar{m}_{H,I(i+1)} = K_{I(i+1)} (\bar{m}_{H,I(i)} \bar{m}_{H,i+1}) \quad (7)$$

where $K_{I(i+1)} = (1 - \sum_{t=1}^{N} \sum_{\substack{j=1 \\ j \ne t}}^{N} m_{t,I(i)} m_{j,i+1})^{-1}$

After all assessments have been aggregated, the combined degrees of belief are generated by assigning back to all individual grades proportionally using the following normalization process:

$$\beta_n = m_{n,I(L)} / (1 - \bar{m}_{H,I(L)})$$

$$\beta_H = \tilde{m}_{n,I(L)} / (1 - \bar{m}_{H,I(L)}), n = 1,\ldots, N \quad (8)$$

β_n generated above is a likelihood to which H_n is assessed. β_H is the unassigned degree of belief representing the extent of incompleteness in the overall assessment.

B. Expected Utility of the ER Approach

Utility is often used to measure people's subjective feelings and preference. In MADA, Expected Utility is used to support evaluation decision [11]. Suppose $u(H_n)$ is the utility of the grade H_n with $u(H_{n+1}) > u(H_n)$ if H_{n+1} is preferred to H_n .

There are two possibilities:

- If all assessments are complete and precise, there will be $\beta_H = 0$ and the expected utility of the attribute y can be used for ranking alternatives, which is calculated by:

$$u(y) = \sum_{n=1}^{N} \beta_n u(h_n) \quad (9)$$

An alternative a is preferred to another alternative b on y if and only if u(y(a))>u(y(b))

- If any assessment for the basic attribute is incomplete, it will be proven that β_H is positive, and the likelihood to which y may be assessed to H_n is not unique and can be anything in the interval $[\beta_n, \beta_n + \beta_H]$.

Yang defined three measures to characterize the assessment for, namely the minimum, maximum and average expected utilities. Suppose H_1 is the least preferred grade having the lowest utility and H_N the most preferred grade having the highest utility. Then the maximum, minimum and average expected utilities on y are given by:

$$\mu_{\max} = \sum_{n=1}^{N-1} \beta_n \mu(H_n) + (\beta_N + \beta_H) \mu(H_N) \quad (10)$$

$$\mu_{\min} = (\beta_1 + \beta_H) \mu(H_1) + \sum_{n=2}^{N} \beta_n \mu(H_n) \quad (11)$$

$$\mu_{avg} = (\mu_{\max}(y) + \mu_{\min}(y)) / 2 \quad (12)$$

The ranking of two alternatives a and b is based on their utility intervals. a is said to be preferred to b on y if and only if $u_{min}(y(a)) > u_{max}(y(b))$; a is said to be indifferent to b if and only if $u_{min}(y(a)) = u_{min}(y(b))$ and $u_{max}(y(a)) > u_{max}(y(b))$.

According to the evaluation framework in Table I and scoring method, the paper selects four E-commerce companies as the assessment samples from four different industries, see Table II. According to literature [11], suppose $\{u(H_1), u(H_2), u(H_3), u(H_4), u(H_5)\}$={0,0.35,0.55,0.85, 1}, we can obtain the first level distributed assessment & utility intervals of the four E-Commerce enterprises, see Table III.

It is clear from Table III, the minimum utility of Company 3 is larger than the maximum utilities of the other three companies. This means that the e-commerce risks of Company 3 is the least. Based on the same principle, the ranking of the four companies is given by

Company 3 $\succ$ Company 1 $\succ$ Company 2 $\succ$ Company 4, where $\succ$ denotes "is preferred to".

IV. CONCLUSION

E-Commerce risks assessment is very important and complex, and the ER approach is one of the best approaches in multi-attribute assessment method. The numerical study of the paper demonstrates the implementation process of the multi-attribute assessment approach. From the result of this paper, we can find out the company's e-commerce risks. The most important thing for the companies is to improve e-commerce risks management capabilities. We all research the e-commerce risk management in the further study.

TABLE II . GENERAL DECISION MATRIX FOR E-COMMERCE COMPANIES

General attribute		Basic attribute	Company 1	Company 2	Company 3	Company 4
Overall performance	ER (ω_1 =0.25)	ER1(ω_{11} =0.2)	H4(0.7)H5(0.2)	H3(0.1)H4(0.8)	H4(0.8)H5(0.2)	H3(0.4)H4(0.6)
		ER2 (ω_{12} =0.3)	H3(0.4)H4(0.6)	H4(0.7)H5(0.3)	H4(0.5)H5(0.4)	H4(1)
		ER3 (ω_{13} =0.2)	H4(0.6)H5(0.3)	H4(0.8)H5(0.1)	H4(0.7)H5(0.3)	H4(0.6)H4(0.4)
		ER4 (ω_{14} =0.3)	H4(0.8)H5(0.2)	H4(0.9)H5(0.1)	H4(0.8)H5(0.1)	H3(0.2)H4(0.8)
	MR (ω_2 =0.25)	MR1(ω_{21} =0.4)	H4(1)	H4(0.5)H5(0.5)	H4(0.6)H5(0.4)	H4(0.7)H5(0.2)
		MR2 (ω_{22} =0.6)	H4(0.4)H5(0.5)	H4(0.8)H5(0.2)	H4(0.9)H5(0.1)	H4(0.7)H5(0.3)
	TR (ω_3 =0.3)	TR1(ω_{31} =0.4)	H3(0.2)H4(0.8)	H3(0.4)H4(0.5)	H4(0.8)H5(0.2)	H4(0.4)H4(0.6)
		TR2 (ω_{32} =0.3)	H3(0.5)H4(0.5)	H3(0.3)H4(0.7)	H4(0.7)H5(0.3)	H3(0.5)H4(0.5)
		TR3 (ω_{33} =0.3)	H3(0.5)H4(0.4)	H4(0.9)H5(0.1)	H4(0.7)H5(0.2)	H3(0.1)H4(0.9)
	IR (ω_4 =0.2)	IR1 (ω_{41} =0.25)	H4(0.8)H5(0.1)	H4(1)	H4(0.9)H5(0.1)	H3(0.3)H4(0.6)
		IR2 (ω_{42} =0.25)	H3(0.3)H4(0.7)	H3(0.2)H4(0.8)	H4(0.8)H5(0.2)	H3(0.4)H4(0.6)
		IR3 (ω_{43} =0.25)	H4(0.6)H5(0.4)	H3(0.2)H4(0.7)	H4(0.5)H5(0.5)	H3(0.7)H4(0.2)
		IR4 (ω_{43} =0.25)	H4(1)	H3(0.5)H4(0.5)	H3(0.1)H4(0.9)	H2(0.1)H3(0.9)

TABLE III. FIRST LEVEL DISTRIBUTED ASSESSMENT & UTILITY INTERVALS OF THE E-COMMERCE COMPANIES

Company	H1	H2	H3	H4	H5	Unknown	Utility Intervals		
							Max	*Min*	*Average*
1	0	0	0.1093	0.7464	0.1011	0.0255	0.8362	0.8107	0.8235
2	0	0	0.1702	0.7410	0.0709	0.0179	0.8123	0.7944	0.8033
3	0	0	0.0025	0.8464	0.1400	0.0111	0.8719	0.8608	0.8664
4	0	0.0101	0.2937	0.5667	0.1186	0.0109	0.7763	0.7653	0.7708

REFERENCES

[1] GreensteinM..“E-commerce:security , risk management andcontr- ol ,”York:McGraw-Hill,2001,pp143.

[2] Luan Xin-xin. “Discussion on Risk Management of E-Commerce Technology,” Inquiry Into Economic Issues,2004,(4),pp96-97.

[3] LiuWei- jiang. “E-commerce Risks and Control Strategy,” Journal of Northeast Normal University(Philosophy and Social Sciences),2005,(11),pp37-41.

[4] Zhu Ya-shu. “Several strategies to achieve the operating system security,” Computer and Digital Engineering,2005,33(1),pp49-52

[5] Zhang Xin. “the Development of Firewall Technology,” Journal of xuzhou Institute of Technology(Social Sciences Edition),2005,(20),pp77-79

[6] Zhao Dong-mei,Zhang Yu-qing,Ma Jian-feng. “fuzzy assessment of the network security risks based on entropy weight coefficient method,” Computer Engineering, ,2004(9),Vol30,No.18,pp21-23.

[7] Lin Bin.,Du Hong-mei.Chen Qing-hua. “Study on the risk assessment based on the probability distribution method,”.Journal of the Academy of Equipment Command & Technology,2006,(8), Vol17,No.4,pp5-9.

[8] Hwang C L, Yoon K. “M ultiple A ttribute Decision Making Methods and Applications: A State of the Art Surveys ,” Springer-Verlag, New York, 1981.

[9] Gabbert P, Brown D E. “Know ledge-Based Computer-Aided Design of Materials Handling Systems,” IEEE Transactions on Systems, Man and Cybernetics, 1989(19): pp:188-196.

[10] Sen P, Yang J B. “Design Decision Making Based upon Multiple Attribute Evaluations and M minimal Preference Information,” Mathematical and Computer Modeling, 1994(20): pp:107-124.

[11] Sen P. “Communicating Preferences in Multiple criteria Decision-making,” the Role of the Designer , 2001, (12): 15-24.

[12] C.L.Huang , K.Yoon, “Multiple Attribute Decision Making Methods and Applications”, A State-of-Art Survey. New York: Springer-Verlag, 1981.

[13] Yager R R, “On the Dempster-Shafer framework and new combination rules”, Information System, 1989, (4),pp93-137.

[14] Yang J B, Singh M G, “An evidential reasoning approach for multiple attribute decision making with uncertainty”, IEEE Transactions on Systems, Man ,and Cybernetics, 1994,24(1),pp1-18.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 289-292

A Model for 10kV Overhead Power Line Communication Channel

Yihe Guo, Zhiyuan Xie, and Yu Wang
Department of Electronic and Communication Engineering
North China Electric Power University, Baoding, China
yihe_guo@163.com

***Abstract*—In order to achieve reliable and high-speed 10kV power line carrier communications, it is necessary to study the characteristics of channel. Based on the multi-conductor transmission line theory, the chain matrix channel model is established. A lumped parameter model of distribution transformer is proposed from impedance measurement in the frequency range 50-500kHz. The terminal voltages and currents of whole network are derived, which reveals the causes of frequency selective fading. The simulation results are compared and analyzed between different signal frequencies and propagation paths, based on a simplified model of distribution network.**

***Index Terms*—power line communication, multi-conductor transmission line, chain matrix, transformer model**

I. INTRODUCTION

The Medium Voltage (MV) distribution line can be used as a communication platform, which is formed by many components with different electric characteristics (e.g., overhead and cable lines). To overcome these problems, recent research efforts are focusing on the investigation of signal propagation and the channel characteristics of the MV network[1].

There are two basic approaches modeling the distribution line communication channel. The first is to assume that the line is so complicated that signal strength can not be accurately computed. Multi-path propagation approaches have been proposed by Philipps[2] and Zimmermann[3]. The parameters for the multi-path model are obtained from measurements of the channel transfer function. This method of statistical channel modeling can not be used to predict the signal propagation without `preliminary` measurement. The second approach is to assume the signal strength can be accurately computed from a transmission line model. In literate [4-5], the channel model is based on two conductor transmission line theory, where the whole network is regarded as a cascade of various two-port networks. So this frequency-domain model works with all the signals reflected from the discontinuities of the network. But this approach can not be applied to MV PLC because the signal propagation in MV is effected by all three conductors and ground.

The classical work in three phase lines model was done by Hardy[6]. The multi-conductor transmission line theory is used to analyze the signal strength on distribution lines. But his result is not applicable to complex distribution network, and the load in his measurement is artificial. The practical MV/LV transformer model for PLC applications is proposed in literate [7], three kinds of measurements are carried out to determinate the parameters of the model. It is obvious that the parameters are varied with the types and capability of the transformer. In order to study the transfer characteristic of the transformer, further measurements are needed.

This paper is organized as follows. First, the chain matrix channel of distribution network is presented in Section Ⅱ. The propagation characteristic of the whole network is analyzed in Section Ⅲ. Section Ⅳ reveals the transmission characteristics of distribution network based on a simplified model. The concluding remark is given in Section Ⅴ.

II. THE MODEL OF DISTRIBUTION NETWORK

For the power line communication, the distribution line should be regard as multi-conductor transmission lines (MTL).

A. MV Overhead Power Line Model

From Fig.1, the 10kV overhead power line communication coupling modes can be divided into phase to phase coupling and phase to ground coupling. If phase to ground coupling is chosen, the signal propagation can also be effected by the other two lines.

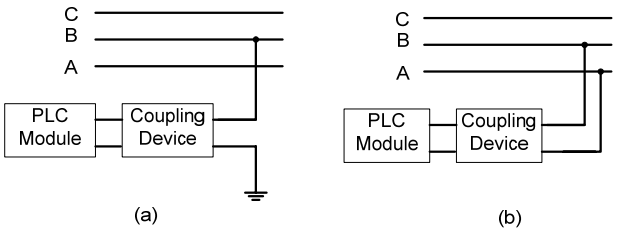


Fig.1.(a)Phase to ground coupling . (b) Phase to phase coupling

The propagation on the MTL is governed by the transmission line equation

$$\frac{d^2}{dz^2}V(z) = ZYV(z)$$
$$\frac{d^2}{dz^2}I(z) = YZI(z) \quad (1)$$

Where $V(z) = [v_1, v_2, v_3]^T$, $I(z) = [i_1, i_2, i_3]^T$ are voltages and currents vectors respectively. The per-unit-length impedance matrix Z and admittance matrix Y are given by

$$Z = R + j\omega L$$
$$Y = G + j\omega C \tag{2}$$

The per-unit-length parameter matrices of resistance R, inductance L, capacitance C, and conductance G can be determined from frequency, conductor and bundle characteristics, line geometry, etc. These line parameters are specified by [3×3] matrices calculated by the Matlab tools of power_lineparam.

In order to decouple the voltages and currents of different conductors, the transfer matrices T_V and T_I are introduced.

$$V(z) = T_V V_m(z)$$
$$I(z) = T_I I_m(z) \tag{3}$$

The 3×3 complex matrices T_V and T_I are said to be similarity transformations between the actual phase line voltages and currents, $V(z)$ and $I(z)$, and the mode voltage and current, $V_m(z)$ and $I_m(z)$, which can be obtained from per-unit lines parameters. The proper T_V and T_I can be calculated, which make $T_V^{-1}ZYT_V$ and $T_I^{-1}YZT_I$ be the diagonal matrices. Then the uncoupled equations are derived as[8]

$$\frac{d^2}{dz^2}V_m(z) = T_V^{-1}ZYT_V V_m(z) = \gamma^2 V_m(z)$$
$$\frac{d^2}{dz^2}I_{\mathrm{m}}(z) = T_I^{-1}YZT_I I_{\mathrm{m}}(z) = \gamma^2 I_{\mathrm{m}}(z) \tag{4}$$

Where γ^2 is the eigenvalue of matrix ZY and YZ.

$$\gamma^2 = \begin{pmatrix} \gamma_1^2 & 0 & 0 \\ 0 & \gamma_2^2 & 0 \\ 0 & 0 & \gamma_3^2 \end{pmatrix} \tag{5}$$

The general solutions to these uncoupled equations are

$$\mathrm{I_m}(z) = e^{-\gamma z}\,\mathrm{I_m}^{+} - e^{\gamma z}\,\mathrm{I_m}^{-}$$
$$V_{\mathrm{m}}(z) = e^{-\gamma z}V_{\mathrm{m}}^{+} - e^{\gamma z}V_{\mathrm{m}}^{-} \tag{6}$$

The actual phase voltage and current can be obtained by similarity transformation

$$\mathrm{I}(z) = T(e^{-\gamma z}\,\mathrm{I_m}^{+} - e^{\gamma z}\,\mathrm{I_m}^{-})$$
$$V(z) = Y^{-1}T_I\gamma(e^{-\gamma z}I_{\mathrm{m}}^{+} + e^{\gamma z}I_{\mathrm{m}}^{-}) \tag{7}$$

The voltages and currents at the two ends of the line can be related with the chain parameter matrix as in (8)

$$\begin{bmatrix} V(l) \\ I(l) \end{bmatrix} = \begin{bmatrix} \phi_{11}(l) & \phi_{21}(l) \\ \phi_{12}(l) & \phi_{22}(l) \end{bmatrix} \begin{bmatrix} V(0) \\ I(0) \end{bmatrix} \tag{8}$$

The chain parameter matrix can be computed as

$$\phi_{11}(l) = \frac{1}{2}Y^{-1}T(e^{\gamma l} + e^{-\gamma l})T^{-1}Y$$
$$\phi_{12}(l) = -\frac{1}{2}Y^{-1}T\gamma(e^{\gamma l} - e^{-\gamma l})T^{-1}$$
$$\phi_{21}(l) = -\frac{1}{2}T(e^{\gamma l} - e^{-\gamma l})\gamma^{-1}T^{-1}Y$$
$$\phi_{22}(l) = \frac{1}{2}T(e^{\gamma l} + e^{-\gamma l})T^{-1} \tag{9}$$

Where l is the length of power line.

Distribution line is not transposed equably, so the similarity transformation matrix can not adopt the uniform ones such as Clarke, Karenbauer. The proper transformation matrix must be based on the actual line parameters.

B. Transformer Model

In literate [7][9][10] the equivalent circuit of transformer in high frequency is proposed. The model is based on the ideal transformer and a block of R, L and C circuit. Some measurements are carried out to determinate parameters.

In addition of the ideal transformer, the follows are taken into account:

1) Winding leakage impedance of each phase.
2) Winding magnetizing impedance of each phase.
3) Winding capacitances including: capacitances between windings and ground, capacitances between windings.

Literate [7] gives the model of the leakage impedance, which is in series with the low voltage load.

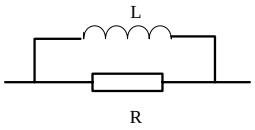


Fig.2. Equivalent circuit of leakage impedance

The measurements of a 160kVA transformer show that the values are

R=2300 Ohm
L=0.176mH

So the magnitude of leakage impedance is beyond 50 Ohm in the range of frequency from 50kHz to 500kHz. The numerous experiments show the characteristics of input impedance of low voltage in China are not consistent with those in Europe and America. The measurements conducted by author show that the input impedance is usually lower than 10 Ohm in the frequency range 50-500kHz, which is accordance with the result of literate[11]. So this input impedance is much lower than the leakage impedance. The MV primary impedance is not virtually depend on the secondary load, which makes the impedance of transformer with load easily determined by winding capacitances. So the simplified transformer model is shown as Fig.3. The nodes ABC correspond to the terminals at the medium voltage.

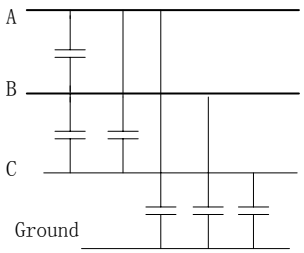


Fig.3. The simplified model of transformer

The admittance matrix of transformer is

$$\begin{pmatrix} jw(C_{AB}+C_{AC}+C_A) & -jwC_{AB} & -jwC_{AC} \\ -jwC_{AB} & jw(C_{AB}+C_{BC}+C_B) & -jwC_{BC} \\ -jwC_{AC} & -jwC_{BC} & jw(C_{AC}+C_{BC}+C_C) \end{pmatrix}$$

C. Branch Line Model

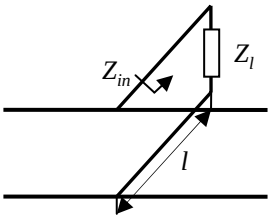


Fig.4. Diagram for the single-branch network

The input impedance of branch line is related to the length of branch line l, the characteristic impedance Z_c, the propagation constant γ and the equivalent impedance Z_l of the transformer connected to the line terminal.

The reflection coefficients of the point z can be expressed as:

$$\Gamma(z) = e^{r(z-L)}(Z_L Y_0 + E)^{-1}(Z_L Y_0 - E)e^{r(z-L)} \quad (10)$$

Where the characteristic admittance matrix Y_0 is :

$$Y_0 = Z^{-1}\gamma \quad (11)$$

The reflection coefficients of the load are given by

$$\Gamma_L = (Z_L Y_0 + E)^{-1}(Z_L Y_0 - E) \quad (12)$$

The input impedance at point z along the line can be obtained as

$$Z_{in}(z) = [E + \Gamma(z)][E - \Gamma(z)]^{-1} Y_0^{-1} \quad (13)$$

Comparing (12) and (13), the expression for input impedance is obtained.

$$Z_{in}(z) = [E + e^{r(z-L)}\Gamma_L e^{r(z-L)}][E - e^{r(z-L)}\Gamma_L e^{r(z-L)}]^{-1} Y_0^{-1} \quad (14)$$

The impedance at input point of the branch line is:

$$Z_{in}(0) = [E + e^{-rL}\Gamma_L e^{-rL)}][E - e^{-rL}\Gamma_L e^{-rL)}]^{-1} Y_0^{-1} \quad (15)$$

The chain matrix of branch line can be described as equation (16).

$$[A_B] = \begin{bmatrix} E & 0 \\ Z_{in}^{-1}(0) & E \end{bmatrix} \quad (16)$$

D. The Whole Model of Distribution Network

The entire power line channel are considered as two parts, with the uniform overhead line being the first, the intrinsic line parameter of which is different; and the equivalent input impedance of branch lines or transformers being the second. They both can be equivalent to the multi-port networks mentioned previously. So the entire distribution line are considered as the cascades of multi-port network.

If the chain matrix of each network is $[A_i]$, the whole matrix can be described as equation (17).

$$[A_a] = \prod_1^n [A_i] \quad (17)$$

III. Analysis of Power Line Propagtaion Characteristic

The 10kV power line carrier communication (PLC) is a technique for transmitting information via distribution line. The power line between the transmitter and receiver can be equivalent to a network in the studying of signal propagation characteristics at the two terminals. According to the practical PLC system, a channel model is illustrated in Fig.5.

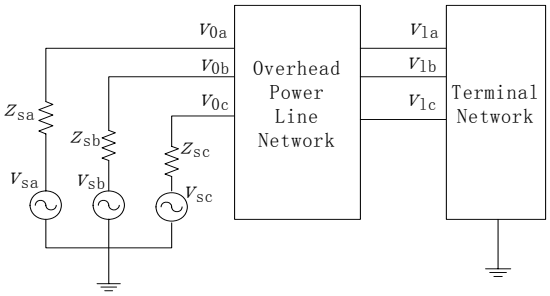


Fig.5. Equivalent circuit of cascaded networks

The transmitter is equivalent to the thevenin circuit, which has voltage sources in series with resistances.

$$Vs = \begin{pmatrix} V_{s1} \\ V_{s2} \\ V_{s3} \end{pmatrix}, \quad Zs = \begin{pmatrix} Z_{s1} & 0 & 0 \\ 0 & Z_{s2} & 0 \\ 0 & 0 & Z_{s3} \end{pmatrix} \quad (18)$$

The vector Vs is the independent voltage of source. Z_S and Z_l contain the effects of the impedances of the source and terminal network.

$$\begin{aligned} V(0) &= V_S - Z_S I(0) \\ V(L) &= Z_L \times I(L) \end{aligned} \quad (19)$$

Incorporate the terminal conditions into the chain matrix characterization given in (20);

$$(\phi_{12}-\phi_{11}Z_s-Z_L\phi_{22}+Z_L\phi_{21}Z_s)I(0)=(Z_L\phi_{21}-\phi_{11})V_s$$
$$I(L)=\phi_{21}V_s+(\phi_{22}-\phi_{21}Z_s)I(0) \quad (20)$$

Then voltage vector and current vector transfer function for the transmission line system can be obtained.

IV. SIMULATION ANALYSIS OF CHANNEL CHARACTERISTICS

A simplified model is presented to reveal the transmission characteristics of distribution network, which is shown in Fig. 6. A is the substation, G、H、D are the distribution transformers at the network terminals. The signal source is located at point B. The coupling mode is phase B to ground.

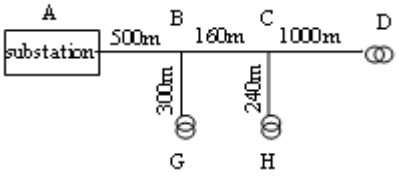


Fig.6. A simplified distribution network model

A. *Model of Each Component*

The models of common components in the distribution network are defined as follows: For the transformer, the capacitances between windings and ground are the same as capacitances between windings. The capacitance is 1000pF for distribution MV/LV transformer. The equivalent capacitance of step-down transformer of 10 kV side in the substation is 22600pF[12]. The power line is equivalent to 3-phases distributed parameters line. The R, L and C line parameters are specified by [3×3] matrices.

B. *Overall Description of Channel Characteristics*

The voltages of D、G on phase B are shown in Fig. 7.

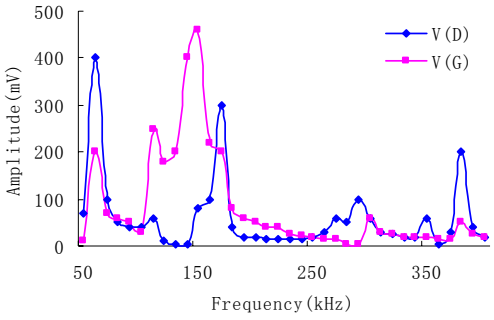


Fig.7. Voltage magnitude of different locations

From the graph we can see that the voltage shows frequency selective fading in the entire frequency band. And this fading is different at each sites. Meanwhile, there are a series of small frequency bands where the attenuation is acceptable for the power line carrier communication.

V. CONCLUSIONS

In this paper, the chain matrix channel model is presented, which includes the power line and distribution transformer. Base on the model, the propagation characteristics of 10kV overhead power line communication channel are analyzed, and the causes of frequency selective fading are discussed. The simulation results show the correctness of analysis. Based on the simulation, several proposals are put forward for the design of communication system.

ACKNOWLEDGMENT

Supported by Youth Foundation of North China Electric Power University through 200711007.

REFERENCES

[1] S.N. Talukdar, J.C. Dangelo, “Uncertainty in distribution PLC attenuation models,” IEEE Transactions on Power Applications and Systems, vol.PAS-99, no. 1, pp. 328-334, 1980.

[2] H. Philipps, “Modelling of powerline communication channels,” International. Symposium on Power Line Communications and its Applications, Lancaster, UK, 1999, pp. 14–21.

[3] M. Zimmermann and K. Dostert, “A Multipath Model for the Power line Channel,” IEEE Trans. Commun, vol.50, no. 4, 2002, pp. 553–59.

[4] S. Galli and T. Banwell, “A novel approach to the modeling of the indoor power line channel-Part II: transfer function and its properties,” IEEE Transactions on Power Delivery, vol. 20, no. 3, pp. 1869 – 1878, 2005.

[5] C.Konate, A.Kosonen, J. Ahola, M.Machmoum, J.F. Diouris, “Power Line Channel Modelling for Industrial Application,” International. Symposium on Power Line Communications and its Applications, Jeju Island, Korea, pp. 76–81,2008.

[6] M.E. Hardy, Sasan Ardalan, J.B. O’Neal, L.J.Gale, K.C.Shuey, “A model for communication signal propagation on three phase distribution lines,” IEEE Transactions on Power Delivery,vol.6,no.3,pp.966-972,1991.

[7] T.Tran-Anh, P.Auriol, T.Tran-Quoc, “High frequency power transformer modeling for Power Line Communication applications ,” Power Systems Conference and Exposition, Atlanta,USA,2006,pp.1069-1074.

[8] R. Paul, Analysis of Multiconductor Transmission Lines. New York:Wily, 1994.

[9] C.Andrieu, E. Dauphant,D. Boss, “A frequency-dependant model for MV/LV transformer," International Conf. on Power Systems Transients (IPST), Budapest, Hungary, June 20-24th, 1999.

[10] Hemminger R C,Gale L J,Amoura F,et al, “ The effect of distribution transformers on distribution line carrier signals,” IEEE Transactions on Power Delivery, vol.2,no.1,pp.36-40,1987.

[11] Zhang Youbing, Cheng Shijie, He Haibo, Xiong Lan, J. Nguimbis, “Modeling of the low-voltage power line used as high frequency carrier communication channel based on experimental results,” Automation of Electric Power Systems,vol.26,no.23,pp.62-66, 2002.

[12] Jiao Shao-hua, Liu Wan-shun, Zheng Wei-wen, et al, “Attenuation analysis of distribution line carrier channels in distribution network,” Automation of Electric Power Systems, vol.24,pp.37-40,Apr.2000.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 293-297

Wind Power Forecasting Based on Time Series and Neural Network

Lingling Li[1,2] , Minghui Wang[2] , Fenfen Zhu[2], and Chengshan Wang*[1]
[1]School of Electrical Engineering and Automation , Tianjin University , Tianjin 300072, China
Email:haohaohao@eyou.com
[2] School of Electrical Engineering and Automation, Hebei University of Technology, Tianjin 300130, China
Email:wmh_16127@yahoo.com.cn
Fenfen337@sina.com
Corresponding author of this paper is marked with *

***Abstract*—The wind farm output power have the characteristics of dynamic, random, large capacity etc, which brought great difficulty for incorporating the wind farm in the bulk power system. In order to rationally regulate the power supply system in large grid connected wind power system and reduce the spinning reserve capacity of the power supply system and operating costs, it is necessary to forecasting the capacity of wind power. For the randomness of the wind farm output, we use the ARMA (q, p) model of time series to forecast wind speed and atmospheric pressure, and using the RBF neural network based on this to forecast wind power. Taking the data of measured wind speed and atmospheric pressure from a wind farm as example, to validate the method described above, and the result show that the method has a certain practicality.**

***Index Terms*—wind power, forecasting, time series, RBF neural network**

I. INTRODUCTION

Wind power is a kind of renewable clean green energy the world's fastest growing at present, and is generally accepted alternative energy technology within a global range. It has become the one of the new main power supply of European and the United States, and also is the important development direction of the renewable energy strategy in China. However, the wind power have the shortcomings of intermittent and volatility etc, the wind power fluctuations need to be balanced through the regulation of standby generators and energy storage system when it was accessed to grid, which is a problem long plagued wind power. If the forecasting of the wind power is available, we can reasonable arrange the operation of conventional generator sets, reduce standby installed capacity of power system, improve power system stability and increase the ability of large power grid to accept the wind power, so it has great significance to implement the forecasting of wind power.

Currently, there are two major steps to complete the short term forecasting of wind farm generated energy: 1, To forecast the wind speed and wind direction in the hight of wind wheel hub of wind generator by using wind speed model, then to calculate the orthogonal wind speed component of the wind speed and wind swept round plane. 2, To calculate the actual output power by using wind generator model.At present, there are two main ways for the latest research: one is based on the physical model; the other is based on the statistical model.

The errors of wind speed forecasting of wind farm mainly related with the forecasting methods, the forecasting cycle and the wind speed characteristics in the forecasting place. In general, the shorter the forecasting cycle and the more relaxed the wind speed change in forecasting site, the forecasting errors will be smaller; on the contrary, the forecasting errors will be larger. Wind speed forecasting can be divided into the short-term wind speed forecasting and the long-term wind speed forecasting. Accurate short-term wind speed forecasting is beneficial to adjust the dispatching plan for the power dispatching departments and reduce the adverse effects to the grid from the wind farm, thus effectively reducing the operating costs and spinning reserve of the power system and beneficial to develop the right electricity exchange program in an open electricity market environment; accurate mid long term wind speed forecasting is beneficial to the planning of wind farm. This article focuses on the short-term forecasting of wind speed.

In this paper, to forecast wind speed by using the method of time series, which the data requirement is low and the cost used to forecast is also low, so it is suitable for the actual operation of businesses. Considering the high degree non-linear relationship showed between the wind speed data and the corresponded generation power, RBF neural network to be used to forecast generation power. And to verify the feasibility and effectiveness of the method in this paper through the experimental data from a wind farm.

Algorithm flow is shown in Figure 1：

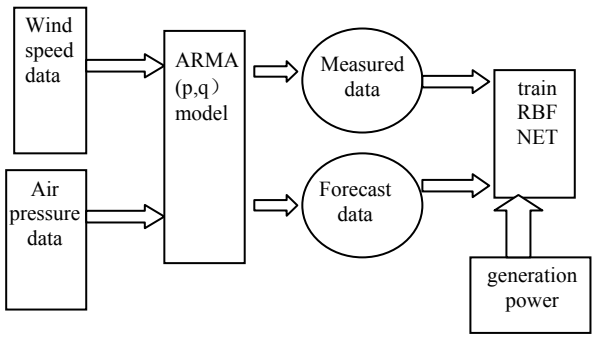


Fig.1 Data flow

II. The Influencing Factors of The Output Power of Wind Farm

The wind power captured by wind turbine can be represent use"(1),"

$$P = C_p S \rho v^3 / 2 . \quad (1)$$

Where: P is output power of Fan, KW; C_p is the power coefficient of Fan; ρ is air density, kg/m^3; S is the area of the wind swept round, m^2; v is the wind speed of the fan windward;

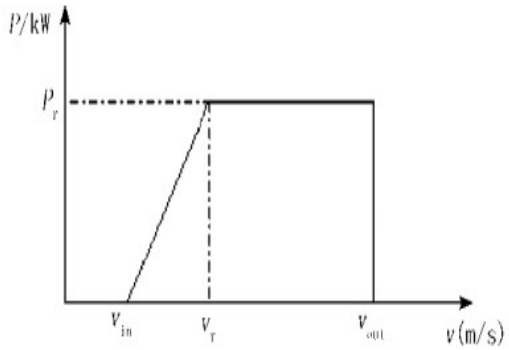


Fig.2 Power curve of a wind power generator

Figure 2 is a power curve of a variable speed wind turbines, in the region of the power curve with more steep, the smaller changes in wind speed will cause the larger changes in power. Figure 3 is a scatter diagram of the measured output power of wind turbines and wind speed, we can see the wind turbines output power has a certain dispersion. This is due to the effects of mechanical turbulence and thermal turbulence, and the spatial distribution of wind speed did not fully comply with the logarithmic wind profile; the other hand, the Yaw device of wind turbines make fans to align the wind direction according to the vane and the wind speed in the height of hub, but there may be some delay, and fans not always being right to the wind direction. This has resulted the different output power with the seemingly similar wind speed

Air density ρ is also an important factor affecting the output power. The power curve of the avariable speed wind turbines with different air density is shown in Figure 4. We can see from the figure, the wind turbines

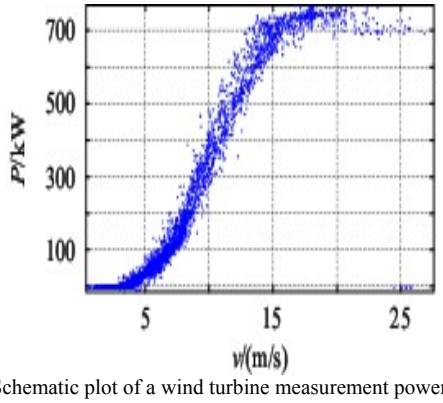


Fig3. Schematic plot of a wind turbine measurement power

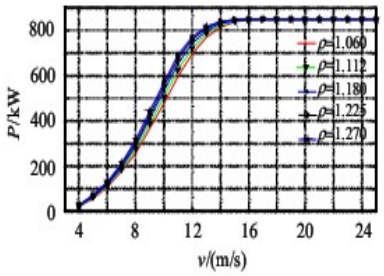


Fig4. A wind turbine power curve under different air density

output power will be larger correspondingly with the air density increasing. Air density closely related with humidity, temperature and pressure. Therefore, the pressure factors need to be considered in wind power prediction.

We can see from the above equation of fan output power, C_p and S are constants can not be changed, and only the air density ρ and the fan windwar wind speed v are variable. So we can know that the main factor affecting power generation of the wind turbine is wind speed and atmospheric pressure.

III. To Forecast Wind Speed Using Time Series

Time series is a modern approach in data processing. At present, the correlation analysis and the cycle burst are the widely used traditional methods. The most fundamental difference between the two is that the parameter model is used to analyse and pretreat dynamic data (time series) for the former method. This method is to set the parameter model that is series model for time dynamic data, and then to obtain the statistical characteristic of dynamic data through this parameter model; and the time series also show the order relationship of the data when it demonstrate the data size.

The wind speed has a very good timing and randomness, and using time series to forecast is more suitable. This paper prepare to forecast the wind speed of wind farm using the Auto Regressive Moving Average model (ARMA) of time series, and this method can forecast only need a single wind speed time series. As follow is the ARMA (p, q) model of wind speed data use"(2),":

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} . \quad (2)$$

Where: p is the autoregressive order of model; q is the moving average order of model; φ, θ are the nonzero undetermined coefficients; ε_t is an independent error term, and the $\{\varepsilon_t\}$ is normal white noise process. It's mean is zero and the variance is δ_ε^2; $\{x_t\}$ is the time series of wind speed.

The most complex problem of ARMA (p, q) model is to determine the order of model. In this paper, using AIC (Akaike information criterion) standard: the minimum

information criterion, while giving the ARMA model technology and the best estimate of parameters, and using the sample with fewer problems. Judging what kind of random process is close to the development process of forecasting goal. Because only when the sample size is large enough, the autocorrelation function of sample can be very close to the autocorrelation function of original time series. For the specific application, making the model order range from low to high within prescribed limits and calculating the AIC value, and finally determining the order maked its value is minimum, which is the right order of the model.

The maximum likelihood estimate of model parameters use"(3),"

$$AIC = (n-d)\log\delta^2 + 2(p+q+2). \quad (3)$$

The least-squares estimate of model parameters use"(4),"

$$AIC = (n-d)\log\delta^2 + 2(p+q+1)\log n. \quad (4)$$

Where: n is the sample number, δ^2 is the fitting residual square sum, d , p and q are parameters. To calculate the values of $\varphi_1, \varphi_2 \cdots \varphi_p$ and $\theta_1, \theta_2 \cdots \theta_q$ by the maximum likelihood estimate and the least-squares estimate after the model order was determined, then the ARMA (p, q) model is determined. Testing the applicability of the model after the model was determined, and testing whether the $\{\varepsilon_t\}$ is white noise, if it is then the model is available, if not then need to re-build model until the test is set up.

In this paper, the raw data are the measured wind speed of a wind farm. The ARMA (4, 3) model modeling by using the method of time series, which is the average

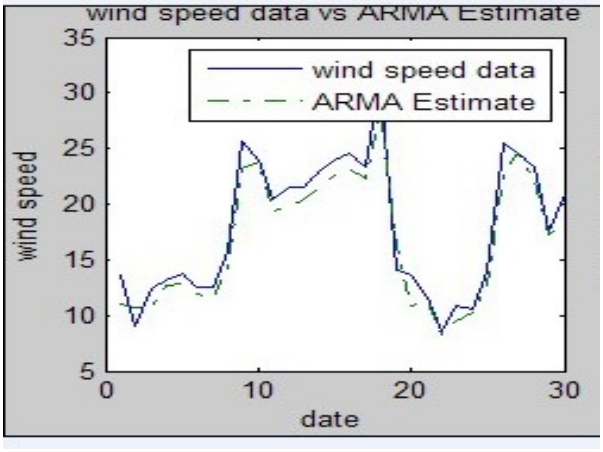


Fig5(a).Results of wind speed forecastin

wind speed, and building a mathematical model using the data of 30 days; the result is shown in Figure.5(a)and autocorrelation of the prediction error in Figure.5(b). The absolute mean error of forecasting was about 24%; the maximum error was 41% emerged in where the wind speed changes most violently. This also indicates that the regularity is stronger and the forecasting accuracy is higher in where the wind speed changes more gentle;

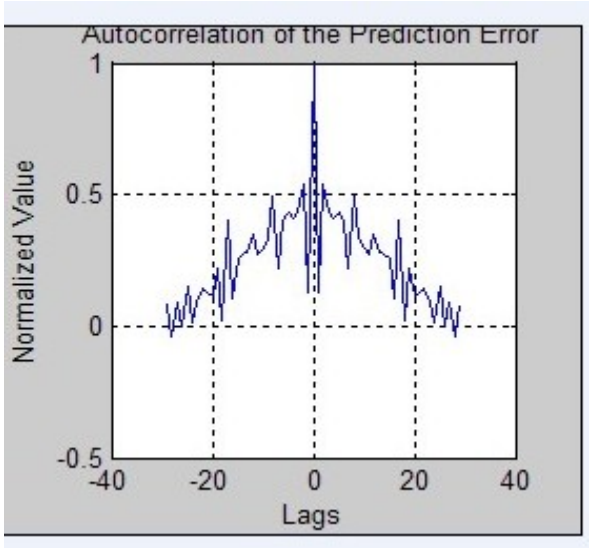


Figure5(b).Autocorrelation of the Prediction Error

while the regularity is weaker and the forecast accuracy is lower in where the wind speed changes more rapid.

IV. EURAL NETWORK FORECASTING OUTPUT POWER OF WIND FARM

The random changes of wind farm output power mainly due to the fluctuations of wind speed and wind direction, while the different fans located in the same wind farm have the almost same wind speed and wind direction. So we can assume all fans in the same wind farm have the same wind speed and wind direction. Using an equivalent wind turbines to simulate the wind farm , and the tail flow coefficient was setted to 0.9,.

The relationship between the wind turbines output power Pw and wind speed v in the height of hub can be approximately expressed by the power curve of wind turbines (Figure 2) or the sub-function use"(5),"

$$P_w = \begin{cases} 0 & v \ge v_{co} \text{or } v \le v_{ci} \\ \dfrac{P_r}{v_r^3 - v_{ci}^3} v^3 - \dfrac{P_r}{v_r^3 - v_{ci}^3} v_{ci}^3 & v_{ci} \le v \le v_r \\ P_r & v \ge v_r \end{cases} . \quad (5)$$

Where: P_r is the rated output power of fan, kW; v is the wind speed in the height of fan wheel hub, m / s; v_{ci} is the cut in wind speed, the automatic device moves the fan into the power grid when the wind speed is higher than this setting value; v_{co} is the cut out wind speed, the fan stops powering and lists from the power system when the wind speed is higher than this value; v_r is the rated wind speed, the fans contribute is rating when the wind speed is higher than or equal to this value but less than the cut out wind speed.

We can see that the strong regularity of the wind speed was further damaged by the wind turbines power curve from the above generation power curve, and got the regularity of wind power is more weaker. Therefore, we introducing the RBF neural network to forecast it in this paper.

Neural network is established based on the basic principles of the biological neural networks, which is a class of adaptive system composed of many simple processing unit called as neurons. Multilayer feedforward neural network can be seen as the non-linear mapping from the input space to output space. It was proved that the forward neural network with one or more hidden layers can approximate any consequent nonlinear function at any degree of accuracy. The study process of neural network is to find a suitable weight vector, and thus be able to approximate function. The size of weight determines the all informations of the neural network. The study process of neural network is a process of amending the weight, which in order to enable the mapping represented by the network be close to the required mapping as much as possible. Back propagation algorithm is a learning method commonly used by multilayer feedforward neural network, which actually is a minimization method with the gradient descent .

Radial Basis Function (RBF) network is proposed by Powell M.J.D in 1985, which is a class of feedforward network taking the function approximation theory as the basic construction. Because the convergence of BP network is slow in function approximation and easy to fall into minimum part, and very incompatibling with the biological context in theory, so in recent years the researchers pay more attention to RBF network. The RBF network is a neural network only containing a hidden layer; the radial basis function is a two-tier network. In the middle layer, it use the radial basis function responsing to part to instead the traditional global responsed excitation function. These good characteristics of RBF shch the simple structure, a fast training process and has nothing to do with the initial value make it was widely applied to the forecast field.

RBF network design include structure design and parameter design. Structure design mainly resolve the problem that how to determine the number of network nodes. Parameter design in general consider the three kinds of parameters including: the data center of basis function , expansion functions and the weight of output node.

According to the method of getting value from data center, the RBF network design methods can be divided into two categories:

The first class method: data centers are selected from the input samples. . Generally speaking, the centers in the density area of samples can be more, and the centers in the sparse area of samples can be less; such as the datas itselves was uniformly distributed, so the centers can be uniformly distributed. In short, the data centers elected should be representative. The expansion constant of r. rradial basis function can be determined by data center, and in order to avoid each radial basis function too sharp or too peace, a selection method is to set all expansion constants of radial basis functions for the use“(6),”

$$\delta = \frac{d_{\max}}{\sqrt{2M}} . \qquad (6)$$

The second class method: the self-organization selection of the data center. It often used a variety of dynamic clustering algorithms to autonomously select the data center, and we need to dynamically adjust the location of data centers in the process of learning. The commonly used method is K-means clustering, and its advantage is that it can determine the expansion constants of the hidden points according to the distance of the each cluster center.

we use the second class method to build radial basis function in this paper, shown in Figure 6:

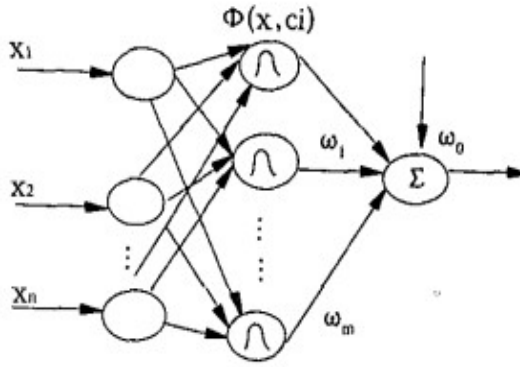


Fig6. Structure of RBF networks

In this picture, x_i is expressed as wind speed and air pressure, and the data output is the output power of wind farm.

The calculated results to the above obtained data using RBF was shown in Figure 7:

We can see from the forecasting maps the absolute

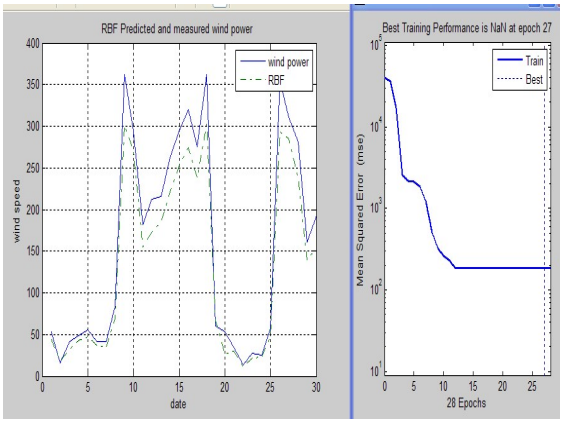


Fig7. Wind power forecasting curve and error band

mean error of forecasting was about 24%; the maximum error was 41% emerged in where the wind speed changes most violently. This also indicates that the regularity is stronger and the forecasting accuracy is higher in where the wind speed changes gentler; while the regularity is weaker and the forecast accuracy is lower in where the wind speed changes more rapid. Its essence lies in that the imprecision of wind speed forecasting and air pressure forecasting directly passed to the wind speed forecasting. If we want to further improve the forecasting

accuracy, we need to find the ways to improve the forecasting accuracy of speed and pressure in wind farm.

V. CONCLUSIONS

The technical requirements(try out) of national grid wind farms accessing grid clearly pointed out the need for forecasting the wind farms power. The analysis of time series and the RBF neural network will be introduced into the wind power forecasting in this article. The forecasting of wind power has a great significance to the construction and operation of wind farm.The study Conclusions of the above mentioned wind power forecasting system have:

Time series model is a dynamic model, which has a very good extension to dynamic data, thereby it could avoid the impact of the directly adding "Window" when we strike the statistical properties of dynamic data. For the random and dynamic of wind speed, the method of time series ARMA reflects a larger advantage.

RBF neural network has a very good non-linear learning ability and a advantage in resolving wind power forecasting.

The Changes in the characteristic watery wind make a great difference between the wind power and conventional power. In actual operation, most wind power has anti-peaking characteristics. As a consequence, the power system in operation must balance the wind power fluctuations with sufficient standby power supply and peak regulation capacity (the spinning reserve) .This not only increases the cost of the entire power system operation, still also bring the hidden dangers to the safe and stable operation in the power grid. In theory, spinning reserve remaining for the wind power are equal with a considerable wind power installed capacity. Therefore, the access to wind power will lead the lower rates of the conventional unit load, and the unit coal consumption to increase.

On the contrary, if the further in-depth study are concentrated on the establishment of this prediction system which are used in the power dispatch center, and the wind electric power are unified into the dispatching plan, then the spinning reserve remaining out of wind power will only be required to meet the forecast error of wind power .As a consequence , the wind power forecast, acted in the city , will increase the rates of the conventional unit load and decrease the unit coal consumption . The wind power forecast will not only improve the economic efficiency, but also conductively make the energy-efficient emission reduction and environmental protection.

In summary, the ARMA-RBF model is not only innovative and pioneering in theory, but also the wind-power forecasting system based on the model has an important practical value and good economic and social benefits.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation (No. 60771069) and and the Hebei Province Natural Science Foundation(No. F2007000115 and No. 602069).

REFERENCES

[1] Xiuyuan Yang, Xiao Yang, Shuyong Chen. Wind Speed and Generated Power Forecasting in WindFarm [J]. Proceedings of CSEE, 2005, 25 (11): 1-5.

[2] Bowden GJ, Barker PR, Shestopal VO, etal. Weibull distribution function and wind power statistics [J]. Wind Engineering, 1983, 7 (2): 85-98.

[3] Tao Sun, Weisheng Wang, Huizhu Dai, etc. Voltage fluctuation and flicker due to wind power [J]. Electric Net technology, 2003, 27 (12): 63-70.

[4] Gaofeng Fan, Weisheng Wang, Chun Liu, Huizhu Dai, etc. based on artificial neural network forecasting of wind electric power. CSEE 2008, 28 (34): 118-123.

[5] Martin Hagan T, Howard Demuth B, Mark Beale H, etc, Neural Network Design [M]. Translation by Kui Dai. Beijing: Mechanical Industry Press, 2002.

[6] Yaonan Wang, Chunshun Sun, Xinran Li. The simulation study of short term wind speed correction with the measured wind speed. Proceeding of the CSEE, 2008, 28 (11): 94-100.

[7] Zhexue Ge, Zhiqiang Sun. Neural network theory and the realization of MATLAB2007. .. Electronics Industry Press, 2007.

[8] Shuzi Yang, Ya Wu, Jianping Xuan, etc. The engineering applications of the time series analysis (second edition). Huazhong University of Science and Technology Press, 2007.

[9] Zongli Jiang. Introduction to Artificial Neural Networks [M]. Beijing: Higher Education Press, 2001.

[10] Florin Iov, Anca Daniela Hansen. Wind Turbine Blockset in Matlab Simulink [D]. Aalborg University. 2004, 3.

[11] J. G. Slootweg, S.W.H.de Haan, H Polinder,W.L.Kling. General model for representing variable speed wind turbines in power system dynamic simulations [C]. IEEE Transactions on Power Systems, 2003, 18 (1).

[12] W. Chu and Z. Ghahramani, Gaussian Processes for Ordinal Regression, J. Machine Learning Research, 2005, Vol. 6, pp1019-1041.

[13] Hyun-Chul Kim and Zoubin Ghahramani， Bayesian Gaussian Process Classification with the EM-EP Algorithm，IEEE Trans PAMI [J], 2006, 28(2): 1948-1959.

[14] P. Sollich, Bayesian methods for support vector machines: Evidence and predictive class probabilities, Machine learning [J]. 2002, Vol.46, pp21-52.

[15] P.Sollich, Probabilistic methods for support vector machines. In Advances in Neural Information Processing Systems 12 (NIPS) [C], S. A. Solla, T. K. Leen and K. -R. Müller, editors, Cambrid ge, MA: MIT Press, pp.349-355, 2000.

[16] D. Q. Zhang, S. C. Chen, and Zhihua Zhou. Learning the kernel parameters in kernel minimum distance classifier. Pattern Recognition[J]. 2006(39): 133-135

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 298-301

Prediction of Electromagnetic Interference to the Switching Operation in Substation

Huijuan Zhang [1], Meng Wu [1], Yanting Wang *[2], Xiaohui Tang [1], and Shitao Wang [1]
[1] School of Electrical Engineering and Automation, Hebei University of Technology, Tianjin 300130, China
Email: zhanghuijuan@hebut.edu.cn , wmxyz135@126.com
[2] School of Management, Hebei University of Technology, Tianjin 300130, China
Email: wangzhanghj@hotmail.com
Corresponding author of this paper is marked with *

***Abstract*—In order to study the electromagnetic compatibility of a power system, authors adopted an effective analysis software and proposed a new prediction method. First, simulated and analyzed the electromagnetic transient when switching the no-load bus in the 500kV substation by the ATP software and proved the simulation results were consistent with the actual results; Second, introduced the Gray prediction theory and used it to predict the next electromagnetic interference of the substation, the result showed its soundness.**

***Index Terms*—switching operation, no-load bus, electromagnetic transient simulation, Gray prediction theory**

I. INTRODUCTION

With the rising of the automation level of power system, more and more electronic equipment are used in the substation, the electromagnetic interference with the substation switching operation is very easy to interfere these secondary electronic equipment or secondary system and affect their normal working. Even affect the safe and reliable operation of the substation and power grid.

In a flash of substation switching operation, it will produce an extremely complex spatial transient magnetic field, space transient electric field and the corresponding transient voltage, transient current, In order to take the necessary protective measures for the electromagnetic interference in the substation, we should make a in-depth study about the transient voltage, transient current and electromagnetic field in the high voltage bus when the switching operation.

II. SIMULATION OF TRANSIENT ELECTROMAGNETIC WHEN SWITCHING THE HIGH VOLTAGE NO-LOAD GENERATOR OF THE SUBSTATION SWITCHING

A. Establish a simulation method by ATP

According to the domestic and foreign experience[1,2,3], the most serious electromagnetic interference arise by the circuit breaker and isolate turning on or off the substation switching. Air-insulated substation is currently the most common types of substation. Therefore, in this paper, major study aims the air-insulated substation, to establish a simulation method of transient electromagnetic field in switching transient process when switching the no-load bus in the 500kV substation by EMTP-ATP and analyze the transient current and transient current.

The figure 1 shows the simplified model of the 500kV substation which only considers a group of bus and lead, with no load on the bus and ignoring the impact of line and structure.

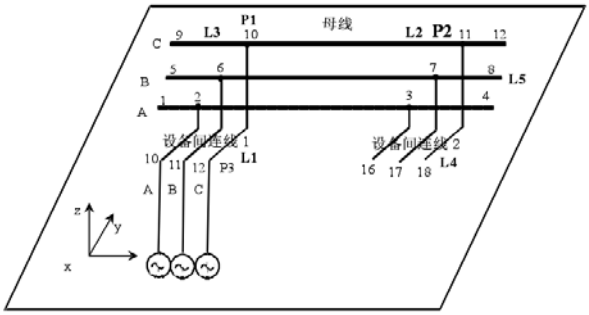


Figure 1. The wiring model of bus with branches in switching transient process

The corresponding simulation method is showed in the figure 2.

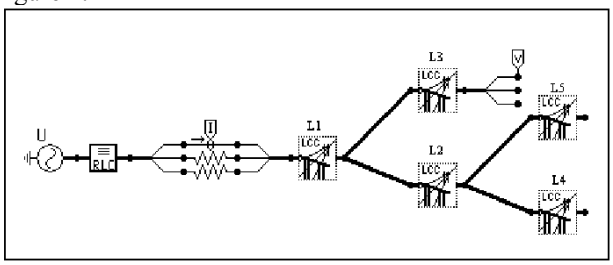


Figure 2. ATP simulation model of bus with branches in isolator switch closing process

B. Get the transient waveforms

After setting up all the parameters, run the ATP simulation software, get the transient voltage waveforms of the related points, showed in the figure 3-5:

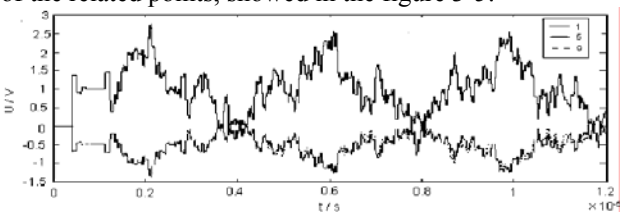


Figure 3. Three-phase voltage waveform of bus in the left end

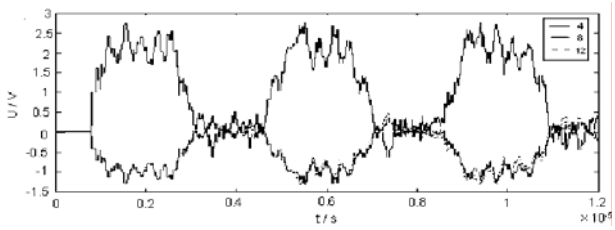


Figure 4. Three-phase voltage waveform of bus in the right end

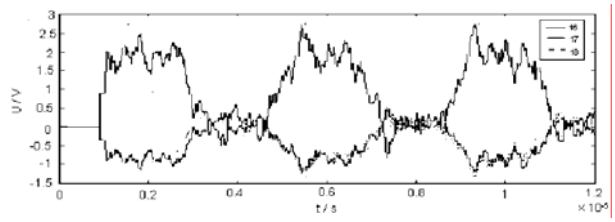


Figure 5. three-phase voltage waveform of line 2 between equipments in the end

The transient current waveforms of P1, P2 and P3 are showed in the figure 6-8:

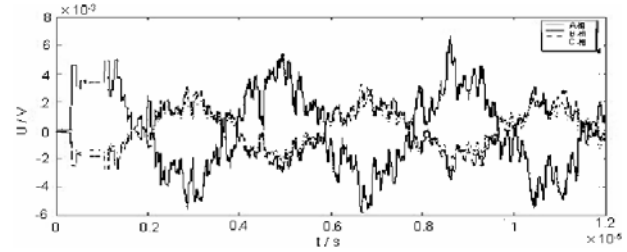


Figure 6. Three-phase current waveform of the three points in P1

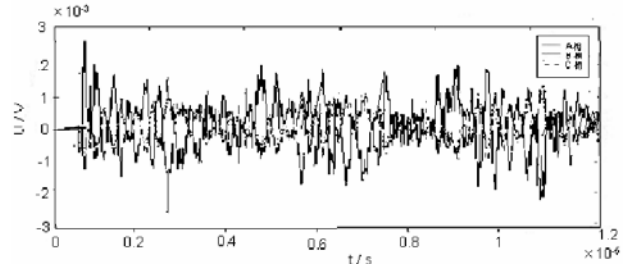


Figure 7. Three-phase current waveform of the three points in P2

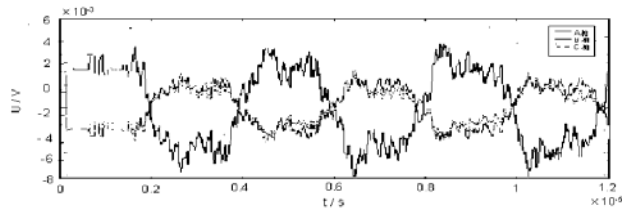


Figure 8. Three-phase current waveform of the three points in P3

Because of the scale of EHV substation is very large, the transient voltage peaks of the different points in different positions have also been differences[3]. Similarly, from the figure we can see that the transient voltage of the bus endpoint is very large in the substation. In different line junction, the voltage there is reflection and refraction make the voltage increasing and dramatic change and the maximum voltage can be amounted to approximately 2.9 times of the input power. In addition, due to the existence of the power branches, making the transient voltage and current waveform more complicated and the burr more intensified than non-branch in the process of switching operation. But the overall trend with and without the branch are same.

Processed voltage wave by Fast Fourier Transform, we could further get the voltage spectrum of points in the substation, showed in the figure 9.

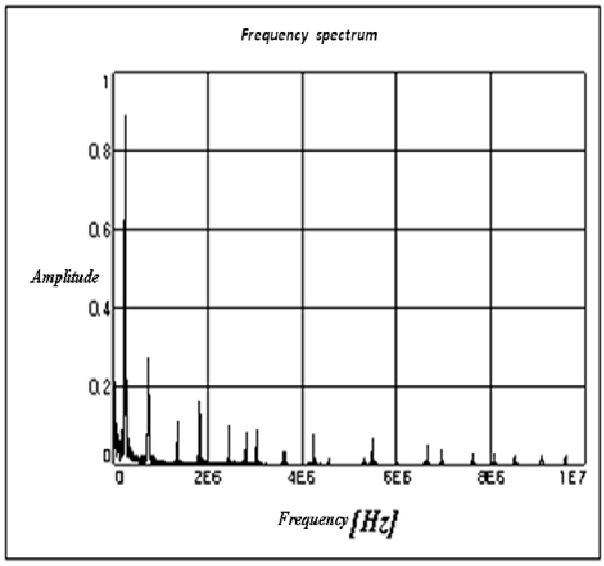


Figure 9. The voltage spectrum of points in A phase

As can be seen from the spectrum, electromagnetic interference spectrum concentrated in the 10MHZ around in the closing of the no-load bus in the substation, it is consistent with the measured data in the literature [4] and [5].

C. Analysis

Comparing the simulation results and the experimental results, the laws of the transient electromagnetic processes were similar with the actual situation. The simulation results could truthfully reflect the actual electromagnetic interference. The simulation is reliable and valid, and it has a high application value for the research of electromagnetic environment in substation.

III. MULTI-VARIABLE GRAY PREDICTION MODEL

A. Improvement the Gray prediction theory

At present, the conventional Gray prediction models include GM(1,1),GM(1,n),MGM(1,n), ect. Because of the unstable changes of value of transient voltage and current with electromagnetic interference caused by switching operation, it must have effects upon the accuracy and validity of models to select them as source data. MGM(1,1) with a simple principle, is not been considered with the selection of original data and background. It may have an negative impact on its precision and availability[6,7]. For the purpose of accuracy promoting, this paper shows several methods to cover the shortage of MGM(1,n) model, and expand its scope of application:

a) Improve the background value by one dimensional search method:

Take any variable $X^{(0)}$ as an example, and

$$X^{(0)} = \{X^{(0)}(P_i) | P_i \in R^+, i = 1,2,...,n\} \quad (1)$$

Define the background value of MGM(1,1) is:

$$Z(P_i) = \beta \times X^{(1)}(P_i) + (1-\beta) \times X^{(1)}(P_i) \quad (2)$$

which 0≤β≤1, change β in proper order, reform matrix B, then we can calculate parameter and prediction value of it. Letbe:

$$q(i_p) = (x^{(0)}(i_p) - \hat{x}^{(0)}(i_p)) / x^{(0)}(i_p) \quad (3)$$

$$q(P_i) = (x^{(0)}(P_i) - \hat{x}^{(0)}(P_i)) / x^{(0)}(P_i) \quad ,$$

P_i are decimals. (4)

which i_p is a positive integer which is less than P_i but close to it.

Then we could get the prediction value:

$$s = \sum (q(i_p))^2 + \sum (q(P_i))^2 \quad (5)$$

Using one dimensional search method, when s is minimized, the value of β is optimum, and it maximizes the precision minimum of error of the prediction model.

b) Improvement the initial conditions:

Albino Gray differential equation is known as:

$$\frac{dX^{(1)}}{dt} + aX^{(1)} = u \quad , \quad (6)$$

its time response function is

$$\hat{x}^{(1)}(k) = C_1 e^{-a(k-1)} + \frac{u}{a} \quad , \quad (7)$$

$$\hat{x}^{(1)}(m) = x^{(1)}(m), (m = 1,2,...,n) \quad (8)$$

as the known condition to count C1, and select the constants C1 which make the average relative error of the model smallest as the optimal initial condition of the MGM (1,1).

The time function of MGM(1,n) as known is

$$X^{(1)}(t) = e^{Am}(X^{(1)}(0) + A^{-1} + C) - A^{-1}B \quad (9)$$

According to the re-modify boundary condition (8),we could get:

$$X^{(1)}(t) = e^{A(t-m)}(X^{(1)}(m) + A^{-1}B) - A^{-1}B \quad (10)$$

So the solution of the modal is:

$$X^{(1)}(k) = e^{A(k-m)}(X^{(1)}(m) + A^{-1}B) - A^{-1}B \, , \quad k = 1,2,...n \quad (11)$$

Here, according to the practical situation, m can be selected from 1,2,…,n.The new formula can be regarded as an amendment or a development of the original one. When m is 1, the two formula are equal.

B. Achieve the Gray predicition model MGM(1,n)

The paper adopts the practical dates of transient common mode voltage in PT end when reclosing three-phase circuit breaker in literature [8] as the input dates.

By analyzing the result of several kinds of transient switching operation test, using the data analysis system of substation transient electromagnetic interference, we take function change the sample data and get another series of dates. They represent results of the parameter analysis using the data analysis system of substation transient electromagnetic interference. The results can reflect the situation of substation transient electromagnetic interference in truth. Table 1 shows the concrete data.

Table 1. The data table

Common mode voltage in PT end					
No.	Time-Domain Characteristic Parameters				
	Up/Down time	Up/Down rate	Duration	Peak	Energy
1	5	59.4	10	1410	56.88
2	10	85.9	20	1500	62.32
3	10	90.6	20	1890	67.62
4	25	102	31	4530	97.07
5	25	216	45	5410	122.48
6	50	241	51	6030	169.66

Then, on the theory bases of improved MGM(1,n), we establish the computation module of Gray prediction models with electromagnetic interference.

Adopting 6 sets of parameters of common-mode voltage, the former four sets of dates are used for modeling, while the latter two sets of data are used to test the accuracy of predictive value.

In table 1, the peak value is the absolute value of the difference between transient positive peak value and negative peak value, with high randomicity and impact resistance, and it is influenced by many factors, so it is not suitable for modeling. Therefore, this article focuses on up (down) time, up (down) rate, duration, energy for model validation.

Using the MGM (1,4) with the initial conditions improved, the fitting of the examples above are shown in table 2, for simplicity, all the error data in the table are relative.

Table 2. The fitted error table when different m

m	Vari-able	No.1 (%)	No.2 (%)	No.3 (%)	No.4 (%)	Average Error (%)
1	X_1	0	22.40	22.58	32.66	19.41
	X_2	0	21.50	10.28	32.85	16.16
	X_3	0	25.26	21.76	23.54	17.59
	X_4	0	21.9	23.03	23.80	17.13
2	X_1	13.25	17.34	16.46	13.22	15.07
	X_2	14.65	18.42	10.66	12.70	14.11
	X_3	16.42	15.64	12.62	10.13	13.70
	X_4	10.88	9.82	14.88	14.20	12.08
3	X_1	17.94	22.29	25.42	20.86	21.63
	X_2	13.64	13.42	18.28	17.07	15.61
	X_3	16.56	24.15	14.67	21.42	19.20
	X_4	20.86	23.08	32.64	26.34	25.73
4	X_1	13.24	16.12	12.13	21.68	15.79
	X_2	20.86	23.22	22.25	16.34	20.67
	X_3	16.67	12.85	18.56	16.42	16.13
	X_4	20.95	13.37	30.28	27.16	22.95

The table above shows the fitting relative error of all prediction models when m=1,2,3,4. According to the table, the fitting accuracy is highest. Therefore, we use the formula with m=2 for prediction which is

$$X^{(1)}(k) = e^{A(k-2)}(X^{(1)}(2) + A^{-1}B) - A^{-1}B \quad (12)$$

Take $X^{(0)}(4)$ as an example, the results are shown in table 3.

Table 3. Forecast results of variables when m=2 in model

Actual value		No.5	No.6	Average relative error(%)
		122.48	169.66	
MGM(1,4)	Predictive value	137.50	161.04	
	Absolute error	15.02	8.62	11.82
	Relative error(%)	12.26	5.08	8.67

Finally, draw the prediction fitting curve with $X^{(0)}(4)$, and the fitting situation of the improved MGM (1,4) as figure 10 shows:

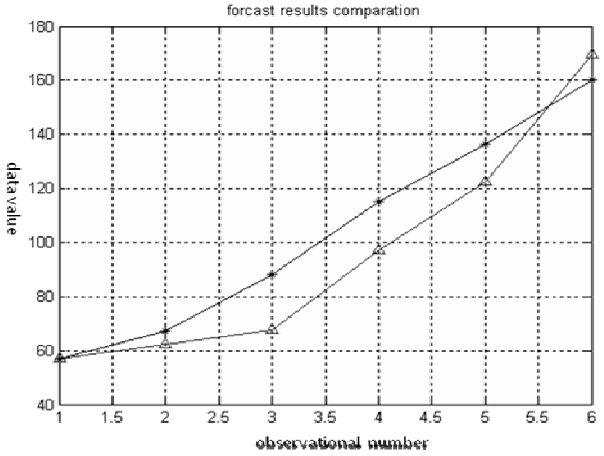


Figure 10. Results comparison between fitted curves of advanced MGM(1,4)and actual value curves

The fitting prediction results of the models above show that, the result is influenced by the poor regularity of the data sources. So there are different between the fitting prediction dates and the original dates.

C. Analysis

Reasons of the tolerance errors between these results:

a) Narrow scope of the study due to the small amount of source data we got. For the repeatability of switching operation is small, and the electromagnetic environment is extremely complex at the same time, we received a limited amount of measured data. There are irresistible factors.

b) We can adopt more effective methods to improve the smoothness of $\{ X^{(0)}(k) \}$, and then develop accuracy of the prediction model. While considering the identity of the raw data and many factors of modeling, these differences are in the allowable range.

It shows that, the prediction measures of this paper with high accuracy, is practical enough to meet the engineering needs.

IIII. CONCLUSION

In this paper, first, we study the electromagnetic environment caused by isolated switching transient operation in both cases of no-load bus with and without branches, and analyze the transient over-voltage and over-current. Compared with the measured results, simulation model has a better accuracy, and it is proved that electromagnetic transient simulation is reliable and effective, and has high practical value for researching and analyzing substation electromagnetic compatibility.

Second, introduce the Gray prediction theory into the study of substation electromagnetic interference, a practical example shows that it is an effective method with rapid forecasting speed, high precision and less dates preparation. It is proved that multivariable Gray prediction theory is applied to model electromagnetic interference prediction analysis is feasible, and has a broad developing space and applied prospect.

APPENDIX AUTHOR

Zhang huijuan: Doctor, Professor, Study for Electrical Theory and New Technology

Del: 13752518191 or 022-60204613,

Postal Address: Post office box 360, Electrical and Electronic Learning Center, School of Electrical Engineering and Automation, Hebei University of Technology, Tianjin 300130, China

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation (No.60771069), Hebei Province Science and Technology Research and Develop Plan Project (No.06547002D-3) and Hebei Education Department Foundation (No.B2001222)

REFERENCES

[1] Thomas D E et al. Induced transient in substation cables: measurement and models. IEEE Trans on PD,1994,9(4):96

[2] Rao M. Mohana, et al. Computation of electromagnetic interference fields in a high voltage gas insulated substation during switching operations. IEEE Int. Symp. Electromagn. Compat, 2003, (2):743~748.

[3] Bo Zhang,Zhibin Zhao,Xiang Cui,etc,Diagnosis of Breaks in Substation's Grounding Grid by Using the Electromagnetic Method, IEEE Transactions on Magnetics,2002，38(2):473-476

[4] Lu tiebing, Study for numerical prediction method of Substation transient electromagnetic environment. [PhD thesis],Baoding, North China Electric Power University,2001

[5] Li ji, Prediction software development for EHV substation switching transient electromagnetic process. [Master thesis], Baoding, North China Electric Power University,2003

[6] Yu jianming, Yan fei, Yang wenyu, et al. Variable weight combination Gray prediction model of long-term electric load, Power System Technology, 2005，29(17)：26~29

[7] Fang rengcun, Power system load forecasting interval, [PhD thesis],Wuhan, Huazhong University of Science and Technology,2008

[8] Wang shuang, Manage and analyze electromagnetic disturbance measurement data of substation switching transient, [Master thesis], Baoding, North China Electric Power University,2003.12

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 302-305

Semantic Retrieval Using Ontology and Document Refinement

Bing Chen, and Xiaoying Tai
College of Information Science and Engineering
Ningbo University
Ningbo, Zhejiang Province, China
dabingdejizhi@126.com, taixiaoying@nbu.edu.cn

***Abstract*—To enhance the retrieval accuracy of information search engine, this paper proposes a information retrieval system based on semantics and document refinement that realized by employing the semantic description and relevance of ontology to the information system. We describe the using of LSI (latent semantic indexing) approach to replace the traditional VSM (vector-space model) approach in detail in the results of sorting process and have a comparative experiment. Various experiments demonstrate the feasibility and effectiveness of our approach, LSI approach proposed in this paper is more effective and able to query the most relevant results in the top of the returnee for semantic retrieval than VSM and about 10.7%~22.2% increase of the performance.**

***Index Terms*—ontology; document refinement; semantic retrieval; LSI; VSM**

I. INTRODUCTION

The major information query approach is keyword matching in existing information retrieval (IR) system, only the query terms appear in the document is likely to be retrieved, so that natural language text with polysemy and synonymy words in the query will be left out that causing lower accuracy. Keyword mismatch is one important reason that affecting the efficiency of IR. Ontology technology can not only rely on the structure and format of information resource, but also the semantic level from the associated knowledge, which can be provided a collection of queries with the same or similar meaning that based on user query into the system to improve the retrieval accuracy of the system that belonging to the traditional approach of query expansion[1,2]. But how to extend the keyword is also a problem that generally believed that the original query words that best reflect the needs of users, so that given to the original query term higher weight, to extend the term given lower weight. Document refinement is an alternative way that the query is known as an information item and the document is amended to seek information match[3]. In this paper, an improved document refinement approach is realized, which solving the keywords mismatch problem in IR system and avoiding the weight configuration problem of query expansion.

The semantic IR approach in this paper has the following characteristics: firstly, ontology Knowledge-based (KB) library built by wordnet[4] (http://wordnet.princeton.edu/) and Chinese wordnet[5] (http://aturstudio.com/wordnet/) is used to document refinement and solved the problem of semantic description of resources; secondly, the link relationship between the instances and documents is established to solve the interrelated resources problem; thirdly, LSI approach is used to replace the traditional SVM approach in detail in the results of sorting process. A.Kontostathis used mathematical methods proved that LSI technology only get the latent semantic structure between words[6]. Various experiments demonstrate the feasibility and effectiveness of our approach, LSI approach proposed in this paper is more effective and able to query the most relevant results in the top of the returnee for semantic retrieval than VSM.

The remainder of the paper is structured as follows. Section II introduces our semantic information retrieval model and semantic retrieval approach. Section III compares LSI with SVM in detail in the results of sorting in our various experiments. Finally, conclusion and future work are given in Section IV.

II. SEMANTIC RETRIEVAL USING ONTOLOGY AND DOCUMENT REFINRMENT

In this paper, text parsing and document refinement used by the ontology knowledge at the first, then we establish the relationship between the instances in KB library and documents in corpus. The users' queries according to submit to the semantic IR system, then the system return a collection of documents that according with the queries condition. The semantic IR model in this paper as below fugure 1. Lucene engine is used to index files of the corpus. LSI approach is used in the sorting of semantic IR system. LSI approach has too high calculated quantities to apply to the large-scale semantic information retrieval system, which it needs to be improved.

A. Text parsing and document refinement

The purpose of the text parsing is extracting the useful information from the unstructured text, while document refinement is according to structure information entities by ontology technology. Before the analytical processing of text content, the entire text will be divided into several small pieces, annotated and extracted terms[7]. The same or similar meaning words are be focus on the specific concept.Generally only the indexed keywords can be

This project is sponsored by the national natural science foundation of China under the grant No. 60472099.

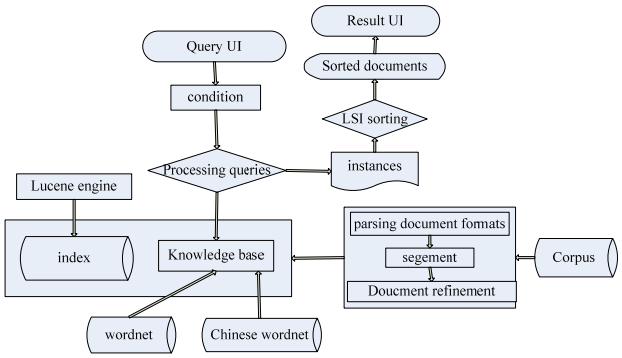


Figure 1. the semantic IR model

retrieved by IR system, so we reorganize document by wordnet and chinese wordnet.

The document refinement algorithm is as follows：

(1) for term K，uses wordnet and chinese wordnet to expanse synonym set{ $s_{k1}, s_{k2}, ..., s_{kn}$ }and hyperonymy set { $h_{k1}, h_{k2}, ..., h_{kn}$ }.

(2) for the term, if $term \in S_k$ or $term \in H_k$, write K and set weight l_k . The local weight setting uses the following strategies: if $term \in S_k$, $l_{\text{term}}(K) = 1$; if $term \in H_k$, and K is n-hyponym of term，set weight as $l_{\text{term}}(K) = \left(\frac{1}{2}\right)^n + \alpha$, α is adjustable parameter.

(3)for the term , if $term \in S_k$ or $term \in H_k$, considered $sim_K(term, K)$ between term and K. Vocabulary semantic similarity subjective appraisal, inconvenience direct calculation, Therefore, to convert it into two words distance measure in wordnet as $dist_K(term, K)$ if K is the synonym of term, $dist_K(term, K) = 1$; if K is n-hyponym of term，$dist_K(term, K) = n$.

(4) The definition of normalized similarity between term and K：

$$sim_K(term, K) = \frac{1}{\beta + dist_K(term, K)}$$, β is adjustable parameter. In order to prevent topic drift，set threshold value T，if $sim_K(term, K)$ greater than T, can used to calculate the final weight $t(K)$.

(5) if K_i Repeatedly used to replace of $term_j (0 \le j \le n)$, shows that K_i make a greater contribution to the document, With the cumulative weighted way to give a higher weight. Finally the calculated weights K_i

$$t(K_i) = \sum_{sim(K_i, \text{term}_j) \ge T}^{i \ne j} l_{\text{term}_j}(K_i) * sim(K_i, term_j)$$ 。

In semantic retrieval processing, the result returns form query interface are instances in KB library, but users want to get the query related to a collection of original document. So that, another function of text parsing and document refinement module is established the mapping between the instances in ontology KB library and the documents in corpus, used to store the corresponding relationship between the instances and documents. With the association table, users can through the query interface returns to get the real linked obtains.

B. Query processing and retrieval ordering

In order to better allow users to express his intention to search, query interface is responsible for the natural language query. Users use natural language queries that summit to IR system, and then the IR system uses ontology and some simple natural language understanding technology to analysis the user's query, extract content which is the users real want to retrieve. IR system constructs these keywords as a query vector, then calculates similarity with the instances matrix, finds the best match instances with the user's query, and returns user the real linked obtain.

In retrieval sorting processing, LSI technology is used in this paper. It is based on the assumption that the word appear-ed in the instances are not random, but exist with a latent semantic structure [8]. In a collection of instances, latent semantic structure can be generated by the use of the singular value decomposition (SVD). Before the SVD decom-position of instances, document refinement approach in this paper for effective feature selection that reduce unwanted characteristics generated in SVD space To improve LSI performance.

LSI technology is based on SVD decomposition. SVD decomposes features matrix of instances A_{mn} to three matrixes: $A_{mn} = TSD^T$, a feature dimension matrix $T_{mr} = (\vec{t_1}, \vec{t_2}, ..., \vec{t_r})$, $\vec{t_1}, \vec{t_2}, ..., \vec{t_r}$ is the left singular vector of A_{mn} and eigenvector of $A_{mn}A_{mn}^T$; Singular value matrix $S_{rr} = diag(\sigma_1, \sigma_2, ..., \sigma_r)$, $\sigma_1, \sigma_2, ..., \sigma_r$ is all the singular values of A_{mn} ; dimension matrix of instances $D_{nr} = (\vec{d_1}, \vec{d_2}, ..., \vec{d_n})$, $\vec{d_1}, \vec{d_2}, ..., \vec{d_n}$ is the right singular vector of A_{mn} and eigenvector of $A_{mn}A_{mn}^T$ [9].

Instances and queries is As a vector formed by the terms in this paper. The instances is expressed as $d_j = (t_1, t_2, ..., t_n)$,

term $t_k(1 \le k \le n)$ is often given a weight w_k. So that the instances can also be expressed as $d_j = (w_1, w_2, ..., w_n)$, and $w_i(1 \le i \le n)$ is the weight value t_i.

This paper uses a improved $tf.idf$ approach to determine the weight value $w_{ik} = tf_{ik}.idf_k$, tf_{ik} has been given a new meaning，Its value is $t(term_i)$ when the document refinement calculated，as follows:

$$tf_{ik} = \sum_{sim(term_i, t_k) \ge T}^{i \ne k} l_{term_i}(t_k) * sim(term_i, t_k)$$

$$= \sum_{sim(term_i, t_k) \ge T}^{i \ne k} \frac{(\frac{1}{2})^n + \alpha}{\beta + dist(term_i, t_k)},$$

and $\alpha = \frac{1}{4}, \beta = \frac{1}{2}$. df_k is the number of t_k appeared in the collection of instances. idf_k is the countdown of df_k, it can be calculated the weight value of terms in each instance. The terms weight value w as a row and the instance d as a column can be formed a large matrix X_{mn}.

Then X_{mn} uses SVD decomposition as $X_{mn} = TSD^T$, only keep top k-max singular value of S_{rr}, get the k-diagonal matrix of S_{rr} and T_{mr} .Finally, inverse of SVD decomposition, get a new matrix：$\widehat{X_{td}} = T_{tk} S_{kk} D_{kd}^T$. D_{kd} is a collapsed dimension matrix of instances which is used to calculate similarity between queries and instances , the aim is to reduce the amount of computation LSI.

The retrieval processing as follows: users submit queries in natural language form to the IR system, query vector is generated by vocabulary frequency $\vec{q}$. $\hat{\vec{q}} = \vec{q}^T T S^{-1}$ is compared with the matrix D_{kd} of instances. $\hat{\vec{q}}$ is the query vector，calculate similarity with each row $\vec{d_j}$ in the matrix D_{kd} ,use the formula as follows:

$$sim(\vec{d_j}, \vec{q}) = \frac{\sum_{i=1}^{k} q_i t_{i,j}}{\sqrt{\sum_{i=1}^{k} (q_i)^2} \sqrt{\sum_{i=1}^{k} (t_{i,j})^2}}$$

It can calculate the relevance between instance $\vec{d_j}$ and query $\vec{q}$. After all instances is calculated similarity with the query , all instances where returns that not less than $S_{\min}$

After the results return, retrieval system through a document-instance table to be able to get a list of documents. Sort list of files to choose, and ultimately returned to the users document a high correlation results.

III. E XPERIMENTAL RESULTS AND ANALYSIS

For corpus, the recall rate is very difficult to statistics. People are more concerned about whether the document retrieval to return to the top surface, as well as the top surface close to each. Therefore, ranking indicator is often used to evaluate information retrieval system performance[10]. When the rate of recall is difficult to calculate, the accuracy has become the main concern of information retrieval systems. p@10(p@20) is the accuracy of the top 10 (20) terms in the query returns. When view the search results, users often want to find the information in the first page of the returns, p@10(p@20) is able to more effectively reflect the real application environments the performance of the system performance[11]. This paper uses Sogou corpus for our various experiments, and the follows experiment results indicate that LSI approach is more adapt to the corpus with higher vocabulary diversity than VSM.

Let $q, r_1, ..., r_m$ as the result of the system returns, $ranking(r_j)$ as the correct position of the D_j for the query q, and the average order value is calculated as follows:

$$AR(r_j) = \frac{1}{m} \sum_{j=1}^{m} ranking(r_j)$$

A query in the search results ranking on average, the smaller of the value the better of the performance.

We are using two types of sorting algorithm model for the random queries to carry out 20 times in 100 before returning results, $AR(r_j)$ ranking indicator of the statistics used two different sorting algorithm:

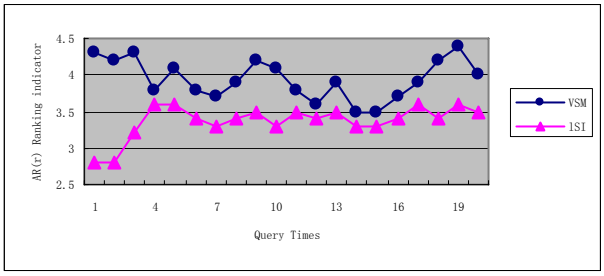


Figure 2. $AR(r_j)$ Ranking indicator of LSI and VSM

In order to reflect the relevant documents with the query q in the search results ranked in the top position of

the close degree of close called the average degree of order. Following formula can be calculated as follows:

$$ART(r_j) = \frac{1}{m}\sum_{j=1}^{m}\frac{j}{ranking(r_j)}$$

If all the relevant documents in the first, then the value of 1.

We are using the two above of sorting algorithm model for the random queries to carry out 20 times in 100 before returning results as the same, $ART(r_j)$ ranking indicator of the statistics used two different sorting algorithm:

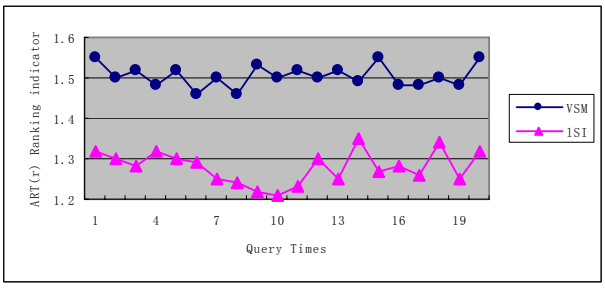


Figure 3. $ART(r_j)$ Ranking indicator of LSI and VSM

We are using the two above of sorting algorithm model for the random queries to carry out 10 times in 100 before returning results as the same, p@10(p@20) of the statistics used two different sorting algorithm ,the results

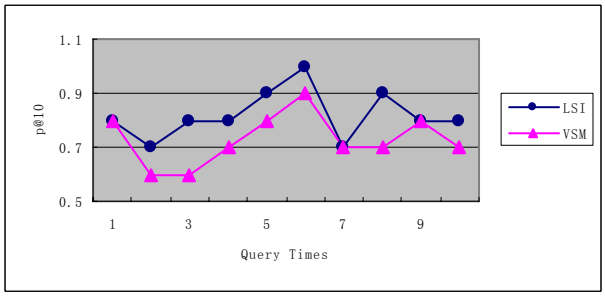


Figure 4. p@10 indicator of LSI and VSM

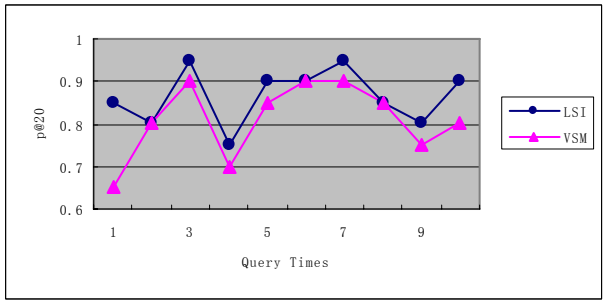


Figure 5. p@20 indicator of LSI and VSM

as the follow figure 5 and figure 6:

From figure 2 to figure 5, the experiment results show that the $AR(r_j)$ Ranking indicator of VSM approach is 4.4, $ART(r_j)$ ranking indicator is 1.50, p@10 value is 0.73, p@20 value is 0.81; When used LSI approach, $AR(r_j)$ Ranking indicator is 3.6, $ART(r_j)$ ranking indicator is 1.25, p@10 value is 0.82, p@20 value is 0.865. LSI approach proposed in this paper is more effective and able to query the most relevant results in the top of the returnee for semantic retrieval than VSM and about 10.7%~22.2% increase of the performance.

IV. SUMMARY

This paper builds knowledge-based ontology liberary by wordnet and Chinese wordnet, adopts document refinement approach in semantic IR system and uses LSI approach to replace the traditional SVM approach in detail in the results of sorting process. But there are two aspects that need to be improved:

(1)Ontology technology: which is used to build KB liberary and document refinement. Ontology directly affects the accuracy of semantic IR system. Ontology is the current hot issue of the semantic web.

(2) LSI sorting performance: which is pretty bad when has a high-dimension of vector. How to reduce the LSI computation is also the current hot issue of the IR research.

REFERENCES

[1] LI Da-gao, CHENG Xian-yi, ZHANG Dong-hui. Query Expansion Algorithm Based on Association Rules and Cluster Algorithm[J]. Computer Engineering ,2009, 33(6): 44-46.

[2] YU Hong, WAn Chang-xuan,. A Semantic Similarity Retrieval Model for XML Documents[J]. Journal of Information ,2007, 26(10): 51-54..

[3] ZHANG Min,S0NG Rui-Hua,MA Shao-Ping. Document Refinement Based on Semantic Query Expansion[J]. Chinese Journal of Computers,2004, 27(10): 1395-1401..

[4] WordNet(a lexical database for English language) homepage: http://wordnet.princeton.edu/.

[5] Chinese WordNet (a lexical database for Chinese language) homepage: http://aturstudio.com/wordnet/.

[6] A. Kontostathis，W. M. Pottenger．A mathematica1 view of latent semantic indexing ： Tracing term cooccurrences[R]. Technical report， LU-CSE-02-006，Dept. of Computer Science and Engineering， Lehigh University，2002．

[7] ZHANG Yu-ming, NAN Kai, MA Yong-zheng. Research on ontology-based information retrieval system models[J].Application Research of Computers,2008,25(8): 2241-2244,2249.

[8] LUO Jing,TU Xin-hui. Chinese Information Retrieval Based on Probabilistic Latent Semantic Analysis[J]. Computer Engineering, 2008, 34(1):199-201.

[9] JI Duo, ZHENG Wei, CAI Dong-feng. Research on Feature optimizati0n in Latent Semantic Indexing[J].Journal of Chinese Information Processing, 2009,23(2):69-76.

[10] Tai Xiao-ying, Bei Yaner. Introduction to Information Retrieval Technology[M]. BeiJing: Science Press,2006,9:21-22.

[11] LIU ting, QIN Bing, ZHANG Yu, Che Wan-xiang. Introduction to Information Retrieval System[M].BeiJing: China Machine Press, 2008, 12:51-52.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 306-309

An Energy Efficient Clustering Algorithm Based on Residual Energy and Concentration Degree in Wireless Sensor Networks

Yuzhong Chen, and Yiping Chen
Department of Computer Science, Fuzhou University, Fuzhou Fujian, P.R. CHINA
yzchen@fzu.edu.cn
Department of Computer Science, Fuzhou University, Fuzhou Fujian, P.R. CHINA
ypchen86@sina.com

***Abstract*—Energy efficiency is the most important issue in research of wireless sensor networks while the routing protocol plays an import role in achieving energy efficiency in WSN. In this paper, the deficiencies of LEACH are analyzed and an improved cluster-based energy efficient routing algorithm called ECHC is proposed. An election weight taking account of the residual energy and concentration of nodes for cluster-head election is introduced in the proposed algorithm. The algorithm also defines a new criterion for non-cluster nodes to choose its cluster head. Experiments are conducted to prove the efficiency of ECHC.**

***Index Terms*—energy efficiency;cluster; node concentration; residual energy**

I. INTRODUCTION

Energy efficiency has been known as the most important issue in research of wireless sensor networks (WSN). So it is of great importance to design an energy efficient routing protocol for WSN. In terms of routing protocol, there are two different solutions from existing works. One is flat routing, each sensor node plays the same role and sends their data to sink node directly which always results in excessive data redundancy and faster energy consumption. The other is hierarchical routing. In hierarchical routing, the entire network is divided into several clusters. Each cluster consists of some source nodes and a cluster head [1]. Sensor nodes, referred as source nodes, can gather information from the monitoring region and send the sensing information to their corresponding cluster head [2]. The cluster head is elected from all the sensor nodes in a cluster according to some criteria, and is responsible for collecting sensing data from source nods. After receiving data from source nodes, the cluster head also performs data aggregation to reduce the data size before sending data to the sink, which further reduces the power expended for data transfer [3]. Clustering-based routing algorithms are more appropriate and efficient than flat routing algorithms in WSN, and the algorithm proposed in this paper is also focused on the improvement of clustering mechanism.

The rest of this paper is organized as follows. Section II gives a brief description of LEACH and its deficiencies. In Section III, the paradigm of ECHC is presented. Section IV describes the details of ECHC. Experiment results and analysis are presented in Section V. Finally, Section VI draws the conclusion.

II. LEACH ALGORITHM AND ITS DEFICIENCIES

Low-Energy Adaptive Clustering Hierarchy (LEACH) is a self-organizing and adaptive clustering protocol proposed by Heinzelman [4] [5] [6] [7]. The operation of LEACH is divided into rounds, where each round begins with a setup phase for cluster formation, followed by a steady-state phase, when data transfers to the sink node occur. Though LEACH uses random election of cluster-heads to achieve load balancing among the sensor nodes [8], LEACH still has some deficiencies which are listed as follows,

- In LEACH, a sensor node is elected as the cluster head according to a distributed probabilistic approach. Non-cluster nodes decide which cluster to join based on the signal strength. This approach insures lower message overhead, but can not guarantee that cluster heads are distributed over the entire network uniformly and the entire network is partitioned into clusters of similar size, and the load imbalance over the cluster heads can result in the reduction of network lifetime.
- LEACH assumes that all nodes are isomorphic, and all nodes have the same amount of energy capacity in each election round which is based on the assumption that being a cluster head results in same energy consumption for every node. Such an assumption is impractical in most application scenarios. Hence, LEACH should be extended to account for node heterogeneity.
- LEACH requires source nodes to send data directly to cluster heads. However, if the cluster head is far away from the source nodes, they might expend excessive energy in communication. Furthermore, LEACH requires cluster heads to send their aggregated data to the sink over a single-hop link. However, single-hop transmission may be quite expensive when the sink is far away from the cluster heads. LEACH also makes an assumption that all sensors have enough power to reach the sink if needed which might be infeasible for energy constrained sensor nodes.

To address the deficiencies listed above, a clustering-based algorithm called ECHC (Energy and Node Concentration Hierarchical Clustering Algorithm) is proposed in this paper. In ECHC, node concentration and

AP-PROC-CS-09CN005

the residual energy of sensor nodes is considered in cluster-head election, and non-cluster node choose its cluster head according to the residual energy of the cluster head and the size of the cluster.

III. ECHC

A. Network Model

Considering a WSN of N sensor nodes are randomly distributed over a region with a size of M*M, in order to simplify the model, ECHC makes some reasonable assumptions,

- The sink node is located outside of the monitoring area with infinite energy.
- Sensor nodes are stationary within a certain period of time after deployment. They also have same computing power and processing power.
- Sensor nodes can dynamically adjust the radio power according to the communication distance, and the communication between nodes is reliable and symmetric

B. Cluster-head Election

Energy consumption of the cluster heads is relatively expensive, so the residual energy of sensor node is the main criteria for the election of cluster head. Furthermore, data aggregation can save considerable energy when the source nodes forming one cluster are distributed in a relatively small region while the sink is far away from the source nodes, because sensor nodes only need much few energy for sending data to the cluster head than sending data directly to the sink when the sink is located at a remote distance [9]. So it is reasonable to infer that the closer source nodes within a cluster, the lower energy they need to consume to send data.

Due to the deduction above, an election weight taking account of the residual energy and the concentration degree of sensor nodes are introduced in ECHC for cluster-head election.

Definition 1: Given a wireless sensor networks of N sensor nodes 1, 2, …., N, $D^r(i)$ is defined to be the concentration degree of node i, namely the number of sensor nodes it can sense during the rth round.

W(j, r) is defined as the election weight of node j in r^{th} round,

$$\alpha = 1/(1+\beta),\ \beta = \frac{E_j^r}{E_j}$$

$$w(j,r) = \alpha\frac{E_j^r}{\overline{E^r}} + (1-\alpha)\frac{D^r(j)}{\frac{N}{K}} \quad (1)$$

Where K is the number of clusters, E_j is the initial energy of node j, $\overline{E^r}$ is the average residual energy of network in r^{th} round. β denotes the residual energy of node j in round r. α is an adaptive factor to adjust the impact of residual energy and concentration degree to election weight. With the reduction of residual energy, α will gradually increases to adapt to the decrease of the number of effective sensor nodes in WSN.

C. Formation of cluster

Definition 2: $S^r(i)$ denotes the size of cluster i. A non-cluster node should take account of this factor to decide which cluster to join in r^{th} round.

In LEACH, the non-cluster nodes only rely on the signal strength of cluster heads to decide which cluster to join. This rule is based on the assumption that sensor nodes are uniformly distributed in the monitoring region which is infeasible in practical scenarios. In most scenarios, if the rule is applied in cluster formation without further consideration, clusters will always differ greatly in the number. The cluster head of a relatively big cluster have to expend much more energy for collecting data from source nodes than the cluster head in a smaller cluster, bringing imbalance of energy consumption cluster heads which in consequence will impact the lifetime of the whole wireless sensor networks. To achieve better load balancing among cluster heads, ECHC introduces a novel criteria for cluster formation. A non-cluster node considers both the residual energy of the cluster head and the size of the cluster when it decides to join a cluster. The formula used to depict the criteria that considers both the residual energy of cluster head and the size of the cluster is listed as follows

$$v_j(i,r) = \lambda E_i^r + (1-\lambda)\left(\frac{N}{K} - S^r(i)\right) \quad (2)$$

Where λ is a factor used to adjust the impact of $S^r(i)$ and E_i^r which can be obtained by experiments. When a sensor node is far away from all cluster heads, it can utilize formula 2 to choose its cluster head which has more residual energy and smaller cluster size than other cluster heads. This criterion will bring a better load balancing to cluster heads and increase the lifetime of WSN, and the performance improvement will be proved in section V.

IV. DESCRIBES THE DETAILS OF ECHC

A. Multi-hop Routing among Clusters

There are mainly two types of transmission in WSN, one is the data transmission between source node and cluster head, the other is the transmission between cluster head and sink. The path for data transmission might be single-hop or multi-hop. In the case of single-hop transmission, if sensor nodes are far away from the cluster heads or the cluster heads are far away from the sink, they have to increase the radio power to transmit sensing data, hasten the energy depletion, which in turn impact the life time of WSN. In the case of multi-hop transmission, cluster heads that are far away from the sink or source nodes that are far away from the cluster heads can send data to the sink over a multi-hop path without increasing radio power. However, source nodes that are close to the

cluster head have to spend much energy to forward data from other source nodes. On the other way, the cluster heads within one hop range of the sink have excessive burden of relaying, therefore when these nodes exhaust their energy, the whole network will fail too. In order to utilize their respective advantages in the best measure and overcome their deficiencies, a hybrid scheme combing single-hop transmission and multi-hop transmission is introduced in ECHC based on a reasonable assumption that the distance between source nodes and cluster heads are always much smaller than that between cluster heads and the sink. In hybrid scheme, source nodes and cluster heads exchange data directly while the cluster heads will send aggregation data to the sink over a multi-hop path if possible. The details of the hybrid scheme is described as follows,

- Step 1: Sink node marks itself as level 1, and broadcast a message to neighbor nodes to construct a routing among clusters. The Message format is (ID, Level) where ID is the sink node ID and the value of level is 1, and the message is sent with "appointment of cluster-heads" messages. The cluster heads one hop away from the sink will receive the message, and then marks itself as level 2 , mark the ID as its father cluster head ID.
- Step 2: Cluster nodes whose level is 2 modify the value of ID and level with its own information and forward the "appointment of cluster-heads" message. Cluster head receiving the message will mark the node ID from which the message is sent as its father cluster ID and use the level value in the message plus 1 as its new level value if its original level value is 0. Then cluster head will forward the message. The routing construction process will end when there is no level 0 cluster head in WSN.
- Step 3: When a cluster head wants to forward the aggregated data to its father cluster head, it should choose the father cluster head with most residual energy as the next hop for data forwarding to avoid the "hotspot" problem. Each cluster head needs to update the information of its own father cluster head's residual energy periodically.

B Construction of cluster and steady communication

1) Setup phase

- Step1: During initialization, the sensor nodes calculate their own concentration degree denoted in definition 1, and mark their own level as level 1.
- Step2: In the initialization phase of network, the sink broadcasts $\overline{E^r}$, the average energy of sensor nodes in "cluster head election" messages. When a node, for example node i, receives the broadcast message, it will compare its own residual energy E_i^r with $\overline{E^r}$ firstly. If $\overline{E^r} \le E_i^r$, node i will calculate the election weight using its own $D^r(i)$ and E_i^r, and then send the weight and its ID to sink node for cluster-head election in a "cluster head election" messages. Otherwise, node i will give up cluster-head election, and choose to join a cluster later.
- Step3: The sink marks its own level as level 1, chooses K sensor nodes with maximum election weight as cluster heads. Sensor nodes that have been chosen to be the cluster heads by the sink will mark themselves as cluster heads. After that, cluster heads will broadcast to neighbor nodes to notify them that it has been elected as a new cluster head, and build multi-hop path to the sink among clusters according to the scheme described in section II.
- Step 4: When a node is elected as a cluster head, it will broadcast "re-join the cluster" messages to other non-cluster sensor nodes. After receiving the broadcast message transmitted over a single-hop or multi-hop path, the non-cluster sensor node need to determine whether it is in the sensing range of the cluster head or not, in other words, whether the broadcast message is transmitted over a single-hop path or multi-hop path. If not, the sensor node calculates the weight about $N^r(i)$ and E_j^r of cluster j to choose to join the cluster with best weight.

2) Steady-state phase

The cluster head will choose the father cluster head that has most energy as the next hop to forward data. This mechanism can avoid the hotspot problem mentioned above.

Right before the beginning of a new round, source nodes should send the information of their residual energy to their cluster heads along with the sensing data for the purpose of cluster reconstruction. According to the residual energy information collected from the source nodes; the cluster heads calculate the average residual energy of their own cluster and then send the result that will be used for cluster formation of the next round to the sink.

V. SIMULATIONS AND ANALYSIS

In this section, several experiments are conducted to evaluate the performance of ECHC. Both ECHC and LEACH are simulated in same scenes for performance comparison. The simulations are carried out with a random network topology with N sensor nodes are randomly distributed in the monitoring area with a size of 100*100 m using MATLAB.

A simple radio model [5] is utilized in simulations, Network lifetime is defined as the number of rounds until the first sensor is run out of energy. For simplicity, we also determine at prior that ten percents of the sensor nodes will be elected as cluster heads. Furthermore simulation will stop when the number of active nodes in WSN is below 10% of N. For performance comparison, we mainly take account of the following three performance parameters, network lifetime, energy efficient and the stability of ECHC.

Fig. 1 shows the number of nodes that deplete their energy as time elapsed. we can find that the round when the first sensor deplete its energy in LEACH is about 750, while in ECHC, the first sensor node is dead in 1250th round, outperform LEACH about 66%. It is obvious that ECHC prolongs network lifetime greatly.

Fig. 2 show the residual energy of the whole network as time elapsed in this simulation scenario. It is obvious that ECHC reserve much more energy than LEACH. The simulation result also prove that ECHC can achieve much better energy efficiency than LEACH, which in turn extends the lifetime of the whole wireless sensor networks.

Fig. 3 shows the simulation result of the third experiment. In this experiment, same simulations are conducted for LEACH and ECHC with the number of nodes varying from 60 to 200, to observe the impact of the number of nodes on network lifetime. We can see that ECHC outperforms LEACH about 65% to 70% in terms of network lifetime in all circumstances. The simulation result also proves the stability of ECHC.

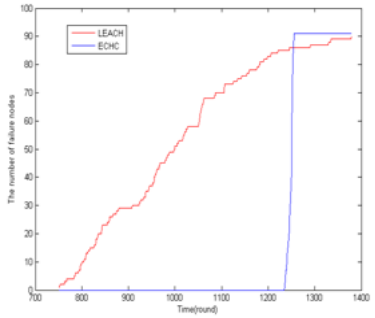


Figure 1. Failure nodes vs Time

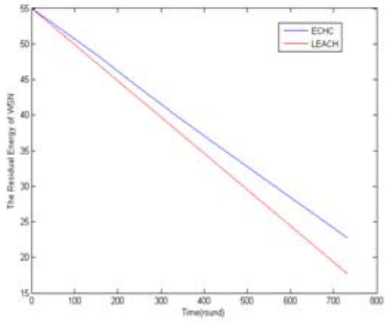


Figure 2. Residual Energy vs. Time

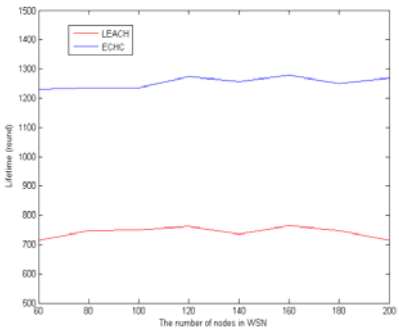


Figure 3 Lifetime vs. Network size

VI. CONCLUSIONS

In this paper, the deficiencies of LEACH are analyzed and an improved cluster-based energy efficient routing algorithm called ECHC is proposed. An election weight taking account of the residual energy and concentration of nodes for cluster-head election is presented in the proposed algorithm. The algorithm also introduces a new criterion for non-cluster nodes to choose its cluster head. Simulation results prove the efficiency of ECHC.

ACKNOWLEDGMENT

This work is supported by the Key Project of Fujian Provincial Natural Science Foundation of China under Grant No.A0820002, the Technology Innovation Platform Project of Fujian Province under Grant No.2009J1007, the Project of Fujian Education Committee under Grant No JA09002, Fujian Province Science Foundation for Youths under Grant No. 2008F3063, the Project of Scientific Research Starting Foundation of Fuzhou University under Grant No. 2008-XQ-24, Special Foundation for Young Scientists of Fuzhou University under Grant No XRC-0827

REFERENCES

[1] E. Fasoloy, M. Rossiy, J. Widmer,M. Zorziy, "In-network Aggregation Techniques for Wireless Sensor Networks: A Survey," *IEEE Wireless Communications*, 2007.

[2] J. Ibriq , I. Mahgoub, "Cluster-based routing in wireless sensor networks: issues and challenges," *Proceedings of the 2004 Symposium on Performance,* 2004.

[3] H. Chen, M. Hiroshi,M. Tadanori, "Adaptive data aggregation scheme in clustered wireless sensor networks," *Computer communication,* vol. 3, pp. 3579-3585, September 2008.

[4] W. R. Heinzelman, A.Chandrakasan, H. Balakrishnan, "Energy-Efficient Communication Protocol for Wireless Microsensor Networks," *Proceedings of the 33rd Hawaii International Conference on System Sciences*, 2000.

[5] S.Ozdemir, Y.Xiao, "Secure data aggregation in wireless sensor networks: A comprehensive overview," *Computer Networks,* vol. 53,pp. 2022-2037, August 2009.

[6] H.S.Lee, K.T.Kim, H.Y.Youn, "A New Cluster Head Selection Scheme for Long Lifetime of Wireless Sensor Networks," *ICCSA 2006*, vol.3983, 2006.

[7] K.O.Younis, S.Fahmy, "HEED: a hybrid,energy-efficient distributedclustering approach for ad hoc sensor networks,", *IEEE Transactions on Mobile Computing,* vol. 3, pp. 366-379, 2004.

[8] V.Gupta, R.Pandey, "Data Fusion and Topology Control in Wireless Sensor Networks," *Proceedings of the 5th conference on Applied electromagnetics, wireless and optical communications*, pp. 135-140 , 2007.

[9] B.Krishnamachari, D.Estrin, S.Wicker, "The Impact of Data Aggregation in Wireless Sensor Networks," *Proceedings of International Workshop on Distributed Event-Based Systems,* 2002

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 310-315

Threshold Visual Cryptography Scheme for Color Images with No Pixel Expansion

Xiaoyu Wu[1], Duncan S. Wong[2], and Qing Li[2]
Department of Computer Science, City University of Hong Kong, Hong Kong, China
[1]Email: xiaoyuwu5@student.cityu.edu.hk
[2]Email: {duncan,itqli}@cityu.edu.hk

***Abstract*—Since the introduction of threshold visual cryptography by Naor and Shamir, there have been many other schemes proposed; some of them support color images with a limited number of color levels while a few others achieve the property of no pixel expansion. However, it is unknown if there is a scheme which can satisfy all the following five commonly desired properties: (1) supporting images of arbitrary number of colors; (2) no pixel expansion; (3) no preprocessing of original images (e.g. dithering or block averaging); (4) supporting *k*-out-of-*n* threshold setting; and (5) a 'tunable' number of color levels in the secret share creation process. In this paper, we answer this question affirmatively by proposing a *k*-out-of-*n* threshold visual cryptography scheme which satisfies all these properties. In particular, our scheme uses a probabilistic technique for achieving no pixel expansion and generically converts any *k*-out-of-*n* threshold visual cryptography scheme for black-and-white images into one which supports color images.**

***Index Terms*—Colored Visual Cryptography, Secret Sharing**

I. Introduction

Visual Cryptography Scheme (VCS), introduced by Naor and Shamir in 1994 [1], is the secret sharing [2] of digitized images. A VCS splits an image into a collection of secret shares which are then printed on transparencies. These shares when separated will reveal no information about the original image (other than the size of it). The image can only be recovered by superimposing a threshold number of shares. This recovery process does not involve any computation. It makes use of the human vision system to perform the pixel-wise OR logical operation on the superimposed pixels of the shares. When the pixels are small enough and packed in high density, the human vision system will average out the colors of surrounding pixels and produce a smoothed mental image in a human's mind.

Early VCS' are mainly focused on black-and-white secret images [3]–[12]. If the original image is not black-and-white, for example, a gray-scale image, dithering [13] is employed to preprocess the original image, that could degrade the image quality. Another issue that is common to most of the previous work is the pixel expansion, which means that each secret share is of size several times bigger than the original image. Two important parameters which govern the quality of reconstructed images are m (pixel expansion rate which represents the loss in resolution from the original image to the shares) and α (the relative difference in weight between the superimposed shares that come from one color level (e.g. black) and another color level (e.g. white)). For image integrity, a good VCS should bring the value of m close to one (i.e. no pixel expansion) and α as large as possible.

In this paper, we propose a new VCS for color images. The scheme has no pixel expansion and allows the original image to have an arbitrary number of colors. The scheme also supports several other desirable properties which are summarized as follows.

1) Supporting images of arbitrary number of colors;
2) No pixel expansion;
3) No preprocessing of original images (e.g. dithering or block averaging);
4) Supporting k-out-of-n threshold setting; and
5) Supporting a 'tunable' number of color levels in the secret sharing process.

In our construction, we generically transform any k-out-of-n threshold VCS for black-and-white images (e.g. [1]) to color images. During the transformation, we use a probabilistic technique for achieving no pixel expansion. In addition, we also allow the user of the VCS to choose the number of colors that the reconstructed image will have. We will see that this 'tunable' feature allows the user to control the quality of the reconstructed image. Based on our experimental results, we believe that this feature can help improve the user friendliness of VCS in practice.

The rest of the paper is organized as follows. In Sec. II, we review some of the related results in VCS; in Sec. III, we introduce some notations which will be used to describe the new VCS; and in Sec. IV, we propose a new threshold color VCS with no pixel expansion. In Sec. V, we propose a grouping method for tuning the number of colors in the reconstructed image and further discuss how the number of colors that could be chosen during the share creation process in Sec. VI. Finally, we provide a quality comparison between our scheme and other related schemes in Sec. VII.

II. Related Work

In [1], Naor and Shamir introduced VCS and proposed several constructions, where the generic one supports k-out-of-n threshold setting for black-and-white images. The scheme does not support images of arbitrary number of colors and the pixel expansion rate is $\log n \cdot 2^{O(k \cdot \log k)}$. Since the introduction of VCS, there have been many other schemes proposed [3]–[12]. In 2004, Adhikari et al.

[5] proposed a VCS which has less pixel expansion than that in [1]. In [14], Yang proposed another one which achieves no pixel expansion. The scheme only supports black-and-white images.

In 2007, Chen et al. extended the results to gray-scale images and proposed a gray-scale VCS [15] with no pixel expansion. However, the scheme does not support the general *k*-out-of-*n* threshold setting. In addition, it also needs to perform block averaging (i.e. preprocessing) on the original image before carrying out the secret sharing. Another gray-scale VCS without pixel expansion was proposed by Chan et. al [16] in 2004. The scheme also needs preprocessing by dithering and adjusting the gray-level of the original image. The general *k*-out-of-*n* threshold setting is not supported either.

For color VCS, [17]–[22], Hou's schemes [17] are considered to be the first set of color VCS'. All the schemes in [17] have the pixel expansion of 4 and do not support the general *k*-out-of-*n* threshold setting and dithering is required for preprocessing the original image. In 2005, Hou and Tu proposed a new color VCS [23]. The scheme also supports *k*-out-of-*n* threshold setting with no pixel expansion. Dithering is still required for preprocessing the original image before secret sharing.

III. PRELIMINARIES

The *k*-out-of-*n* threshold color VCS proposed in this paper supports original images of any number of color levels. Without loss of generality, we herewith assume that the color of the original image is represented by the conventional 24-bit color primitives, *R* (red), *G* (green) and *B* (blue), each has 256 levels (i.e. 8-bits), that is, for each pixel of the original image, the color quality is represented by three bytes of values; and each byte specifies the intensity of the corresponding color primitive: *R*, *G* and *B*. In the following, we introduce some notations which will be used in the rest of the paper.

Consider a generic *k*-out-of-*n* threshold VCS for black-and-white images (e.g. [1]), suppose the pixel expansion rater is *m*, we use an $n \times m$ Boolean Matrix *S* (below) to denote the secret sharing process, namely *n* rows corresponding to *n* shares and *m* columns corresponding to the "colors" (1 for black; 0 for white/transparent) of the *m* pixels of each share.

$$S = \begin{bmatrix} S_{0,0} & S_{0,1} & \cdots & S_{0,m-1} \\ \vdots & & & \\ S_{n-1,0} & S_{n-1,1} & \cdots & S_{n-1,m-1} \end{bmatrix}$$

where $S_{i,j} \in \{0, 1\}$. Depends on the actual black-and-white VCS, the pixel expansion rate *m* varies. For example, if the Naor-Shamir *k*-out-of-*n* VCS [1] is applied, $m = \log n \cdot 2^{O(k \cdot \log k)}$.

A *k*-out-of-*n* black-and-white VCS typically consists of two $n \times m$ Boolean Matrices B^0 and B^1, which correspond to the white and black of a pixel in the original image, respectively. Let
C_b = {matrices obtained by permuting the columns of B^b} where $b = 0, 1$. The secret sharing of the original image is performed pixel-by-pixel. For each pixel in the original image, if the color is white (resp. black), one $n \times m$ Boolean Matrix in C_0 (resp. C_1) is randomly pixel and used for creating the *n* shares. Due to the page limitation, we refer readers to [24] for details.

For *3*-out-of-*4* black-and-white VCS, below is an example of the base Boolean Matrices:

$$B^0 = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \quad B^1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Here, the pixel expansion is 6. In our scheme described in the next section, we will see how to convert this scheme to a threshold *color* VCS with no pixel expansion (i.e. the pixel expansion rate would be 1).

IV. A NEW *K*-OUT-OF-*N* COLOR VCS

We now describe the VCS which supports all the five properties listed in Sec. I. Along with the scheme description, we use Lena image for illustration. The scheme consists of four steps:

1. Histogram Generation: Three histograms representing the intensity distribution of *R*, *G* and *B* color primitives of the original image are first generated. In the histogram for *R* (resp. *G* or *B*), the horizontal axis represents the intensity of *R* (resp. *G* or *B*) ranging from 0 to 255; and the vertical axis represents the number of pixels of each intensity value. Fig. 1 shows the original Lena image and Fig. 2 shows the *R*, *G* and *B* components of it. Fig. 3 shows the three histograms generated in this step.

2. Color Quality Determination: As we can see in the previous step, each color component has 256 levels of intensity. In our scheme, we can let the user choose the number of intensity levels that the reconstructed secret image will have. In this step, the user is to determine this intensity level with the purpose of maximizing the quality of the reconstructed image. (Please refer to Sec. V and VI for details of choosing the number of levels.) Let *N* be the number of levels of reconstructed image where $N=N_R \times N_G \times N_B$ (N_X denotes the number of levels of the primitive $X \in \{R, G, B\}$). Suppose user would like to show a 64-color reconstructed image, then he may choose the number of levels for *R*, *G* and *B* as follows: 64 (*N*) =4 (N_R $\times 4$ (N_G) $\times 4$ (N_B). Remark: the color levels of the reconstructed image for *R*, *G* and *B* do not need to be the same.

3. Grouping: For each color primitive $X \in \{R, G, B\}$, we create N_X groups on the histogram of *X*. To do so, we specify the boundary color intensity between every pair of adjacent groups as $K_0, \cdots, K_{N_X-2}$. In other words, we divide the histogram of *X* into N_X regions, i.e. [0, K_0), [K_0, K_1), $\cdots$, [K_{N_X-2}, 255]. The principle of dividing is to make each of these N_X regions to have the same size in area. So for each group, there will be an equal of pixels in the image that fall into each group/region. In Sec. V, we further discuss the reasons behind choosing this approach

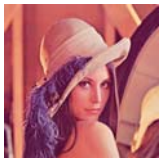


Fig. 1. The Original Lena Image.

Fig. 2. The RGB Component Images of Lena

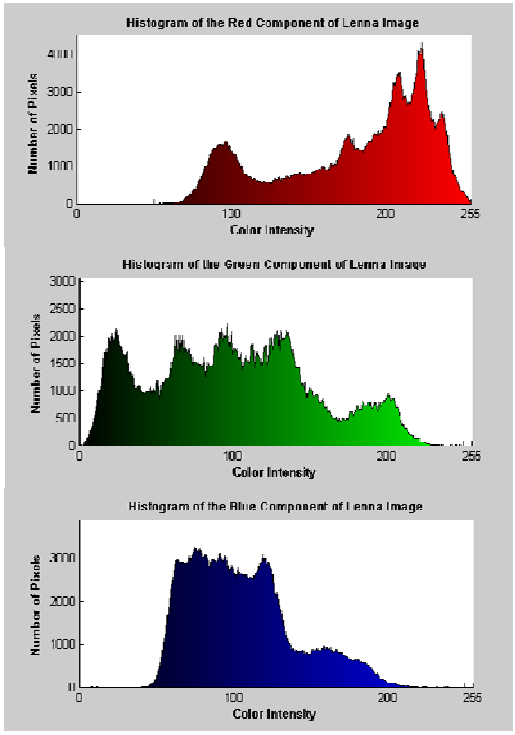


Fig. 3. Histograms of the RGB Component Images

for grouping. Fig. 4 shows the histograms after grouping. Here we use different color intensities to represent different groups.

4. Share Creation: The last step is to create the secret shares. To do so, we apply the following method to each of the primitive color independently. First, take a *k*-out-of-*n* black-and-white VCS, for example, the Naor-Shamir VCS [1]. For the base Boolean Matrices B^b (for b=0,1), denote them as:

$$B^b=[B^b_0, B^b_1, \cdots, B^b_{m-1}]$$

That is, B^b_i denotes column i ($0 \leqslant i \leqslant m-1$) of B^b. Second, for each color primitive $X \in \{R, G, B\}$, we carry out the following steps for each of the pixels in the original image. For each pixel,

1) suppose the color intensity of the pixel with respect to color primitive X falls into the k-th group (where $0 \leqslant k \leqslant N_X - 1$). We compute a probability value $P=k/(N_X-1)$

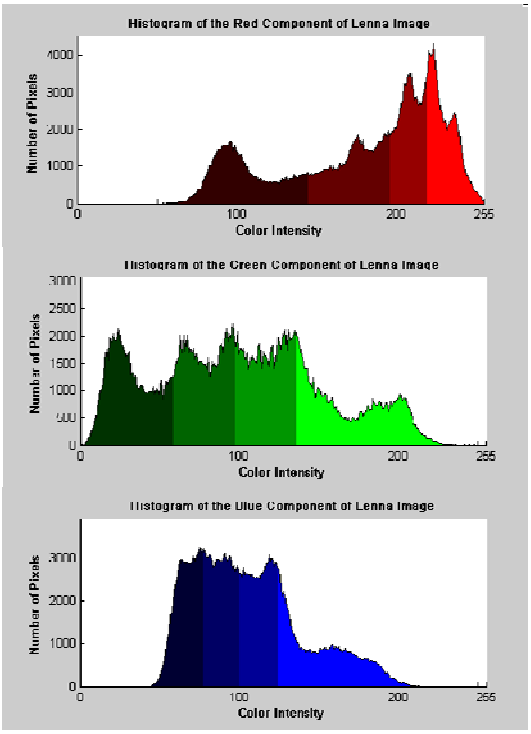


Fig. 4. Histograms Illustrating the 4 × 4 × 4 Color Levels

which determines the likelihood of going through one of the following steps.

2) With the probability P, we carry out the following two steps:

• We look into B^0 and randomly pick a column, for example, B^0_j where $0 \leqslant j \leqslant m-1$.

• Consider B^0_j as an n-bit vector. For the first bit, we assign the black color (i.e. 0 color intensity) if the bit is 1, otherwise we assign red color (i.e. 255 color intensity). This continues until we have assigned colors to this pixel for all the n shares.

3) With the probability 1-P, we carry out similar steps to the above, but change B^0 to B^1.

Determined by the value of P, in Table I, we summarize the probability distribution of B^0 and B^1 for individual groups. Since the columns in B^0 (resp. B^1) are chosen uniformly at random, the chance of picking any particular column in B^0 (resp. B^1) for Group k will be $k/((N_X-1)m)$ (resp. $(1-k/(N_X-1))/m$).

TABLE I
THE CHANCE OF USING B^0 OR B^1 FOR INDIVIDUAL GROUPS DURING SHARE CREATION ($X \in \{R,G,B\}$)

	Probability of using B^0	Probability of using B^1
Group 0	0	1
⋮	⋮	⋮
Group k	$k/(N_X-1)$	$1-k/(N_X-1)$
⋮	⋮	⋮
Group (N_X-1)	1	0

Finally, we superimpose the i-th R share with the i-th G share as well as the i-th B share, for i=1,⋯, n, to form the final i-th share which consists of the corresponding R, G, B components.

A. Example

Suppose we want to create a 64-level 3-out-of-4 set of secret shares such that each color component has four groups, that is $N_R = N_G = N_B = 4$. After dividing the R, G, B components of the original image into four groups (i.e. Grouping), we carry out the Share Creation by first computing the probability value for each group. Table II shows the probability distribution of the individual columns of the base Boolean Matrices B_0 and B_1 of the 3-out-of-4 black-and-white VCS described in Sec. III.

Column 2 of Table II specifies the pixel color of the four secret shares when the i-th column of B_0 or B_1 is chosen (0 represents coloring the pixel of the corresponding secret share to the primitive color; 1 represents coloring it to the black color). Column 3 to 6 indicate the probabilities of choosing the i-th column if the pixel color in the original image is in one of these four groups.

V. Grouping Methods - 'Tunable' Number of Color Levels in Reconstructed Images

In the scheme description above, after generating the histogram for each primitive color and deciding the number of groups, we do the grouping (step 3) by dividing the histogram into several regions so that each region contains the same number of pixels. There are many other ways of doing the grouping. As an example, one can evenly divide the histogram into N_X regions for each primitive color $X \in \{R,G,B\}$ based on the color intensities, that is, making N_X equal-width regions: [0, 255/ N_X), [255/ N_X, 255× 2/ N_X),⋯,[255 × (N_X − 1)/ N_X, 255]. Fig. 5 shows the histogram of R, G and B of Lena image when $N_R = N_G = N_B = 4$. From the figure, we can see that most of the pixels of Lena image on primitive color B are grouped into the second region. After superimposition, these pixels will show no difference in the reconstructed image. From this example, one can see that our approach of making the groups with same number of pixels per region rather than with the equal color-intensity interval is for improving the color level difference among pixels which have different color levels in the original image.

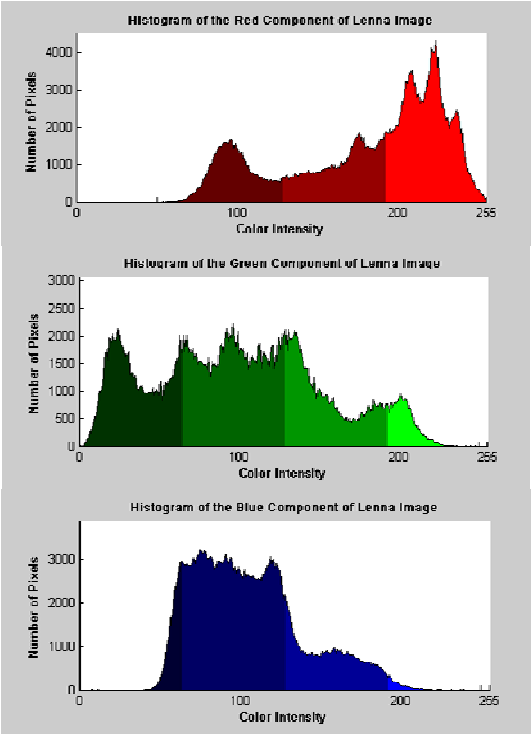


Fig. 5. Histograms with 4 Equal-Width Regions

TABLE II
The Chance of Using Any One Particular Columns of B^0 or B^1 for different Groups During Share Creation

The possibility of choosing the i-th column of matrix M	Shares	Group 0	Group 1	Group 2	Group 3
i=1, M=B^0	0000	0	1/18	1/9	1/6
i=2, M=B^0	0000	0	1/18	1/9	1/6
i=3, M=B^0	1110	0	1/18	1/9	1/6
i=4, M=B^0	1101	0	1/18	1/9	1/6
i=5, M=B^0	1011	0	1/18	1/9	1/6
i=6, M=B^0	0111	0	1/18	1/9	1/6
i=1, M=B^1	1000	1/6	1/9	1/18	0
i=2, M=B^1	0100	1/6	1/9	1/18	0
i=3, M=B^1	0010	1/6	1/9	1/18	0
i=4, M=B^1	0001	1/6	1/9	1/18	0
i=5, M=B^1	1111	1/6	1/9	1/18	0
i=6, M=B^1	1111	1/6	1/9	1/18	0

VI. Discussions on Determing the Number of Color Levels

In this section, we discuss how to choose the number of color levels (i.e. N=N_R×N_G×N_B) in the reconstructed secret image. The scheme supports an arbitrary number of color levels which affects the quality of reconstructed image in a significant way. We observe that the number of color levels to be chosen depends on the number of colors that the original image has. We first classify the original images into two categories: in category 1, the number of levels on a particular primitive color is small, for example, less than 4; and in category 2, the number is large, say at least 4. For images in category 1, if the original image on a particular primitive color $X \in \{R,G,B\}$ is $OriginalN_X$, since $OriginalN_X$ in this case is small, there is no need to try with different color levels. Hence we should set N_X to $OriginalN_X$. Images that fall into this category could be some text and logos. Fig. 6 shows the original image and reconstructed image of Logo of Mac with color levels set to 2 (N_R)×2 (N_G)×2 (N_B).

For images in category 2, one may try the color level N_X from a small value, say 2 or 4, to the "full" level

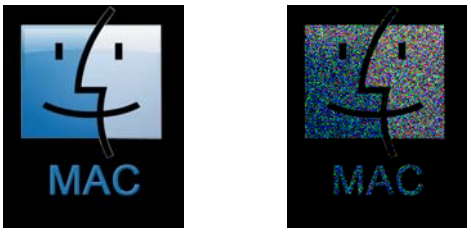


Fig. 6. The Original and Reconstructed Image of Mac Logo (Color Levels: 8 = 2×2×2)

OriginalN_X. Based on our experimental results shown below, we observe that for photos or color cartoon images with large number of color levels, trying these three values (namely 2, 4 or *Original*N_X) for the value of N_X can already attain one of the best results in the reconstructed image.

In Fig. 9 we can see that the reconstructed image of "Alice in the Wonderland" (Fig. 8) with $2\times2\times2$ levels has the sharpest image but limited number of colors while the "full" level version, i.e. $256\times256\times256$, which has the same number of colors as the original one, looks blurry. Image with $4\times4\times4$ shows the best result with abundant colors and clear figure. Fig. 7 and Fig. 11 show similar results as that in Fig. 9. Fig. 13 shows the reconstructed image of Gray scale (21 levels) (original image: Fig. 12) with 4 levels and "full" level (i.e. 21 levels). Note that in this image, the values of *R*, *G* and *B* components are the same for every pixel. We can see that "full" level version gives better result as the gradual change in the gray intensity can better be chosen than that of the 4 level.

VII. Quality Comparison

In this section, we compare our VCS with eight other schemes: Naor-Shamir (NS in short) (the first VCS), Hou (the first colored VCS), Yang (a probabilistic method for gray images), Chan et al. (no pixel expansion for gray images), Hou-Tu (HT in short) (no pixel expansion for color images), Shyu, Chen et al. (multiple-level and no pixel expansion for gray images) and Yang-Chen (YC in short) (a probabilistic method for color images). Table III shows the comparison of the schemes, where *C* is the number of colors of the original image, *m* is the pixel expansion rate and $m_1 = \log n \cdot 2^{O(k \cdot \log k)}$. For the column "Level", it indicates whether there is a limitation on the number of color levels of the original image. Compared with other VCS, the new scheme proposed in this paper supports color images and lets users choose the number of color levels in the reconstructed images based on their desired image quality. Besides, the original image does not need to preprocess such as dithering, which would degrade the quality of reconstructed images. Furthermore, the scheme does not have pixel expansion. When

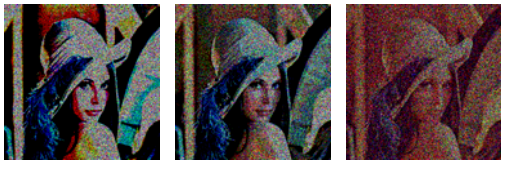


Fig. 7. Reconstructed image of Lena with $2\times2\times2$, $4\times4\times4$, $N\times N\times N$ levels

Fig. 8. The Original Image of Alice

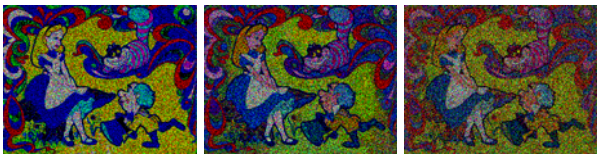


Fig. 9. Reconstructed image of Alice with $2\times2\times2$, $4\times4\times4$, $N\times N\times N$ levels

Fig. 10. The Original Image of F22-Raptor

Fig. 11. Reconstructed image of F22-Raptor with $2\times2\times2$, $4\times4\times4$, $N\times N\times N$ levels

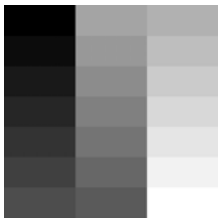


Fig. 12. The Original Image of Gray (21-Level)

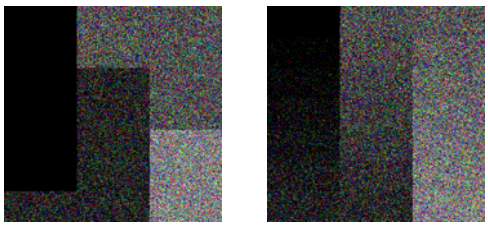


Fig. 13. The Reconstructed Image of Gray (21-Level) with 4 and 21 Levels

TABLE III
COMPARISON

	Colored	Expansion rate	General	Level	tunable
NS [1]	B/W	m_1	√	2 levels	×
Hou [17]	√	4	k=n	8 levels	×
Yang [14]	B/W	1	√	2 levels	×
Chan [16]	Gray	1	k=n=2	2 levels	×
HT [23]	√	1	√	2 levels	×
Shyu [19]	√	$\log_2 C \cdot m$	√	no	×
Chen [15]	Gray	1	k=n	no	×
YC [22]	√	3	√	no	×
Our scheme	√	1	√	no	√

compared with Chen et al.'s [15], we can see that their scheme only supports gray scale images. Also, their scheme does share creation based on the average color intensity of a block of pixels, and the number of color levels of the reconstructed image depends on the block size. The larger the block is, the more levels the reconstructed image has. However, more levels also mean that the more pixels would have the color intensity averaged, thus the quality is degraded.

Yang-Chen's [22] scheme also uses the probabilistic method and support color images. It has a fixed expansion rate 3. The scheme does not support tunable color levels for the reconstructed images.

VIII. CONCLUSION

In this paper, we proposed a new VCS which satisfies the following five properties: (1) supporting images of arbitrary number of colors; (2) no pixel expansion; (3) no preprocessing of original images (e.g. dithering or block averaging); (4) supporting k-out-of-n threshold setting; and (5) a 'tunable' number of color levels in the secret share creation process. According to our experimental results, we show that besides for the first time achieving all these desirable properties, our scheme can provide one of the best reconstructed images in quality due to the 'tunable' feature in the secret share creation step.

REFERENCES

[1] M. Naor and A. Shamir, "Visual cryptography," in *Advances in Cryptology - EUROCRYPT '94*, 1994, pp. 1–12, Lecture Notes in Computer Science, Vol. 950.

[2] A. Shamir, "How to share a secret," *Communications of the ACM*, vol. 22, pp. 612–613, Nov. 1979.

[3] M. Naor and A. Shamir, "Visual cryptography II: Improving the contrast via the cover base," in *International Workshop on Security Protocols*, 1996, pp. 197–202, Lecture Notes in Computer Science, Vol. 1189.

[4] G. Ateniese, C. Blundo, A. D. Santis, and D. R. Stinson, "Visual cryptography for general access structures," *Inf. Comput.*, vol. 129, no. 2, pp. 86–106, 1996.

[5] A. Adhikari, T. K. Dutta, and B. Roy, "A new black and white visual cryptographic scheme for general access structures," in *Progress in Cryptology - INDOCRYPT 2004*, 2004, pp. 399–413, Lecture Notes in Computer Science, Vol. 3348.

[6] C. Blundo, A. D. Santis, and D. R. Stinson, "On the contrast in visual cryptography schemes," *J. Cryptology*, vol. 12, no. 4, pp. 261–289, 1999.

[7] G. Ateniese, C. Blundo, A. D. Santis, and D. R. Stinson, "Extended capabilities for visual cryptography," *Theor. Comput. Sci.*, vol. 250, no. 1-2, pp. 143–161, 2001.

[8] P. A. Eisen and D. R. Stinson, "Threshold visual cryptography schemes with specified whiteness levels of reconstructed pixels," *Des. Codes Cryptography*, vol. 25, no. 1, pp. 15–61, 2002.

[9] C. Blundo, P. D'Arco, A. D. Santis, and D. R. Stinson, "Contrast optimal threshold visual cryptography schemes," *SIAM J. Discrete Math.*, vol. 16, no. 2, pp. 224–261, 2003.

[10] C. Blundo and A. D. Santis, "Visual cryptography schemes with perfect reconstruction of black pixels," *Computers and Graphics*, vol. 22, no. 4, pp. 449–455, August 1998.

[11] S. Cimato, A. D. Santis, A. L. Ferrara, and B. Masucci, "Ideal contrast visual cryptography schemes with reversing," *Information Processing Letters*, vol. 93, no. 4, pp. 199 – 206, February 2005.

[12] M. Uno and M. Kano, "Visual secret sharing schemes with cyclic access structure for many images," in *Information Security and Cryptology (ICISC 2008)*, 2009, pp. 84–97, Lecture Notes in Computer Science, Vol. 5461.

[13] R. W. Floyd and L. Steinberg, "An adaptive algorithm for spatial grey scale," in *Proc. the Society of Information Display*, vol. 17, 1976, pp. 75–77.

[14] C. N. Yang, "New visual secret sharing schemes using probabilistic method," *Pattern Recognition Letters*, vol. 25, no. 4, pp. 481–494, March 2004.

[15] Y. F. Chen, Y. K. Chan, C. C. Huang, M. H. Tsai, and Y. P. Chu, "A multiple-level visual secret-sharing scheme without image size expansion," *Information Sciences*, vol. 177, no. 21, pp. 4696–4710, November 2007.

[16] C. S. Chan, Y. W. Liao, and J.-C. Chuang, "Visual secret sharing techniques for gray-level image without pixel expansion technology," *Journal of Information, Technology and Society*, vol. 95, no. 1, 2004.

[17] Y. C. Hou, "Visual cryptography for color images," *Pattern Recognition*, vol. 36, pp. 1619–1629, 2003.

[18] R. Lukac and K. N. Plataniotis, "A cost-effective encryption scheme for color images," *Real-Time Imaging*, vol. 11, pp. 454–464, 2005.

[19] S. J. Shyu, "Efficient visual secret sharing scheme for color images," *Pattern Recognition*, vol. 39, no. 5, pp. 866–880, 2006.

[20] C. N. Yang and T. S. Chen, "Reduce shadow size in aspect ratio invariant visual secret sharing schemes using a square block-wise operation," *Pattern Recognition*, vol. 39, no. 7, pp. 1300–1314, 2006.

[21] S. Cimato, R. D. Prisco, and A. D. Santis, "Colored visual cryptography without color darkening," *Theoretical Computer Science*, vol. 374, pp. 261–276, 2007.

[22] C. N. Yang and T. S. Chen, "Colored visual cryptography scheme based on additive color mixing," *Pattern Recognition*, vol. 41, no. 10, pp. 3114–3129, 2008.

[23] Y. C. Hou and S. F. Tu, "A visual cryptographic technique for chromatic images using multi-pixel encoding method," *Journal of Research and Practice in Information Technology*, vol. 37, no. 2, pp. 179–191, May 2005.

[24] B. W. Leung, F. Y. Ng, and D. S. Wong, "On the security of a visual cryptography scheme for color images," *Pattern Recognition*, vol. 42, no. 5, pp. 929–940, May 2009.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 316-320

Delegation Management in Service Oriented Decentralized Access Control Model

Houxiang Wang, Ruofei Han, Xiaopei Jing, and Hong Yang
Information and Electric College, Naval University of Engineering, Wuhan, China
hrf_402@sina.com

***Abstract*—Net-Centric Environment (NCE) and Service Oriented Architecture (SOA) are new emerging concept that influence development of information systems. They bring in high risk as well as high sharing. In order to ensure the security of the future military systems, a service oriented decentralized access control (SODAC) model is proposed in previous work. The management to delegation is investigated in further for SODAC model. The entire delegation mechanism and process is discussed, and the three most concerned issues in delegation are well fulfilled with the proposed mechanism.**

***Index Terms*—net-centric, service oriented, access control, delegation management, trust negotiation**

I. Introduction

Delegation is an important issue in access control area. It is an activity that one party hands over its authority to another to accomplish certain task [1]. So the authorization is controlled by the delegation provider according to his own policy, which is out of the range of centralized access control from some point of view. However, there are still several issues that are concerned widely in behalf of the delegation provider, including that:

- Minimum authority delegation. Only the necessary authority for the target task should be delegated, so that the risk of authority leakage can be minimized.
- Time limit to delegation period. Once the target task is accomplished, the delegation should be banished. Otherwise, it may be misused by the delegation requester.
- Control on re-delegation. The permission for delegation requester to re-delegate the authority he achieved from delegation must be constrained in behalf of the original delegation provider, or the authority leakage will be inevitable.

Many researches have been done on these problems [2], but few of them take all the three into consideration. With the development of network and information technology, nowadays systems are almost designed according to service-oriented architecture (SOA), both in commercial or military applications. The latest architecture product, DoD Architecture Framework version 1.5 (DoDAF v1.5), proposed by department of defense of USA, has absorbed the spirit of SOA to guide development of military systems in Net-Centric Warfare (NCW) [3].

Though the Net-Centric Environment (NCE) enables high resource sharing and collaboration between information systems, it also increases the possibility of malicious attack and information leakage. On the other hand, resources and control authorities are both highly distributed in this open environment. To better manage authentication and authorization in this environment, we have proposed a service oriented decentralized access control (SODAC) model in our previous work [4]. In this research, we will further investigate the delegation management in SODAC model. The entire delegation mechanism and concerned issues will be discussed in it.

II. Related Work

A. SODAC model

SODAC model is proposed to meet security requirements of the complicated changefully NCW. It simplifies all the elements involved in the access process into "Entity" and "Service". Service stands for independent function or working process; Entity stands for all the other participating elements in the service. Since these participating elements can be service invoker in one working process, or service provider in another, we named them "Entity" uniformly. For better management, the uniform "Entity" can be divided into several subclasses according to practical applications. Entities are described with "Attribute"; and Services are constrained with "Policy", which are both subclass of "Entity". The structure of SODAC model for military application is shown in Figure 1.

Service composition is a mechanism to issue high level functions that take existing services as a part. SODAC model absorbs spirit from workflow management in task-and-role based access control to manage service composition. There are many benefits to introduce composition in. Firstly, it takes full advantage of previous achievements, so that no repeated development is needed, and much more complicated application would be feasible. Secondly, it facilitates developers to compose new functions rapidly according to user's requirement. Lastly, it shares system load with others, which will also make full use of resources all over NCE. However, the composing service may not be stored locally, even be issued by another domain. In order to protect the information involved in the collaboration, the collaborating parties should come to an agreement on accessing rules through trust negotiation, and delegate corresponding authorities to the other side, which is the most concerning issue in this research.

AP-PROC-CS-09CN005

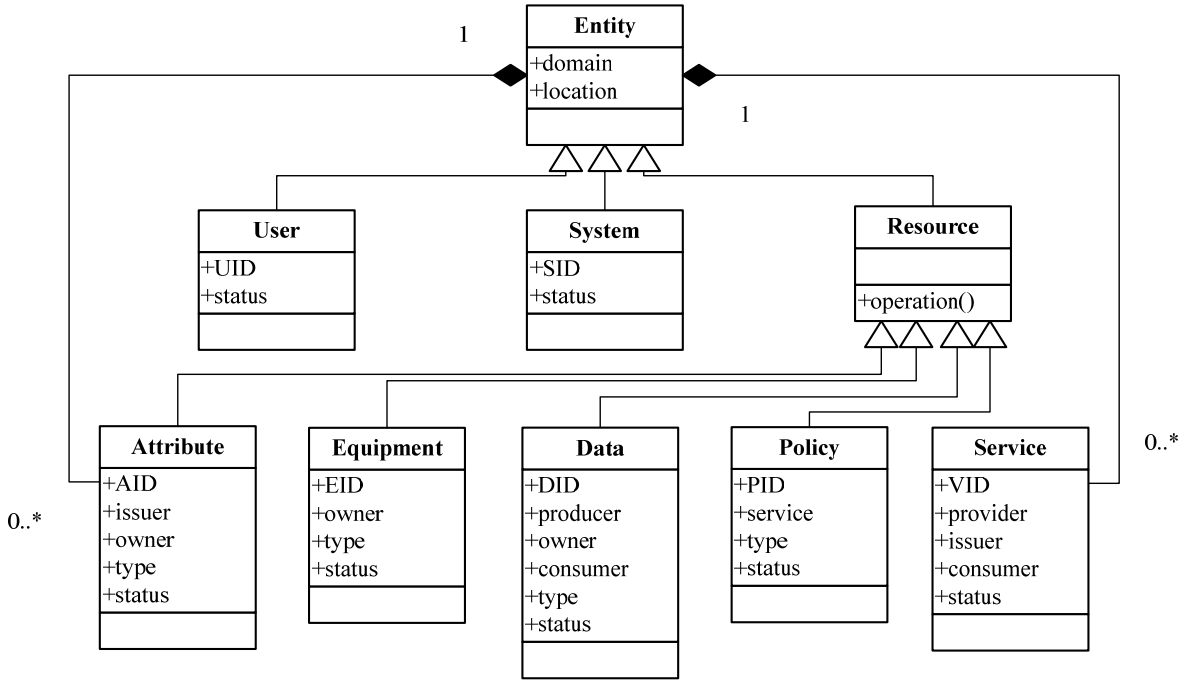


Figure 1. Structure of SODAC model

B. Membership-based negotiation

If the two sides involved in access or delegation process are unfamiliar, they have to negotiate in behalf of each side before collaboration. In our previous work, we have proposed a membership-based access control mechanism for trust negotiation [5].

The membership-based access control classifies external subjects into several groups according to predicted accessing manner. Each group is associated with several particular services, which is provided to members of the group, and with particular policy that is used to constrain subjects to join into the group. Unlike the role in role-based access control (RBAC), which is used to describe responsibility and ability of users in internal working process [6], group is used to describe the level of service that can be provided to its member, and the corresponding conditions on them. So, group is also a set of services and a set of policies. Also, the group set is constructed with various relationships, as shown in Figure 2. The circle stands for policy, and the space surrounded by circle stands for group. If a user wants to become member of a group, he has to satisfy the policy, as the dashed line going across the circle from outside to inside.

When an accessing subject invoked a service whose corresponding membership he does not have, a negotiation route will be produced to guide him joining into the group according to his attributes and group structure. Take the "$G_T \subset G1 \cup G2$" in Figure 2. *c*) for example. Suppose that a subject wants to invoke a service provided to members of G_T, and he is an entirely new customer who even does not have membership of *G1* and *G2*, then the negotiation route will guide him to be member of *G1* and *G2* in the first step as required in *P1* and *P2*, and to be member of G_T in the second step to satisfy requirement of P_T. So the policies, *P1* and *P2*, works just as the prepositive policy of P_T. If the subject can not pass *P1* and *P2*, P_T will not be disclosed to him. However, if he passes *P1* and *P2* but does not satisfy P_T, he can be authorized to invoking services provided to members of *G1* and *G2*, which may be restricted issue of the target service.

However, the negotiation for delegation may be different in some way, but the main mechanism is also applicable.

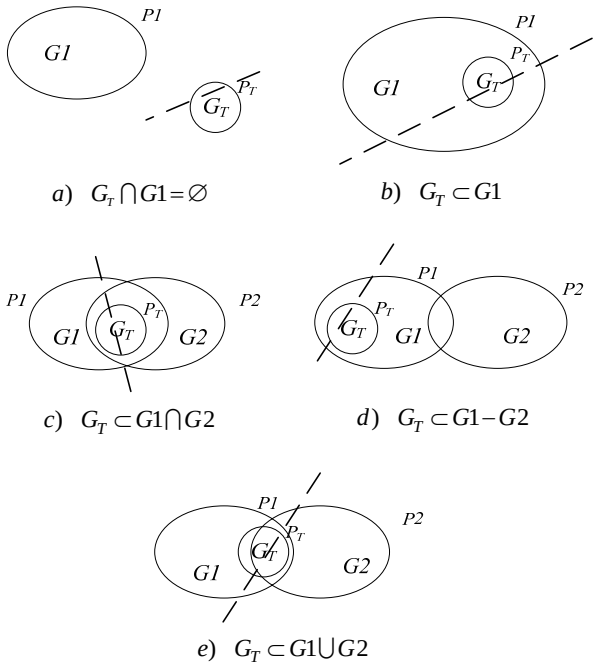


Figure 2. Membership described with group relationship

III. DELEGATION MANAGEMENT IN SODAC

In SODAC model, accessing to resources must be performed though invoking the corresponding service. However, the authorization to service invocation lies on authentication of the invoking entity's attributes according to service policy. So the delegation of authority is turned to be transfer of attribute certificate in the end. Membership can also be described as an authority attribute, which can be introduced here to simplify the delegation authorization.

A. Virtual attribute

In order to constrain the ability of the delegation requester, we introduce in a new kind of attribute, named *virtual attribute*.

Definition 1 (*Virtual Attribute*): a derivative attribute issued by delegation provider to delegation requester to indicate delegated authority.

As shown in Figure 3, the virtual attribute is just a certificate assertion without attribute value, and the provided accessing service is just an invoking interface to that of original service. Accessing services of virtual attribute are also issued by delegation provider according to the authority delegated to delegation requester, which is just a subset of the provider's authority.

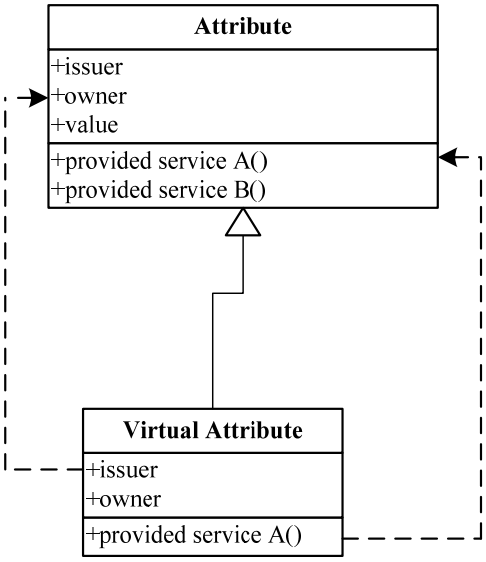


Figure 3. Relationship between Attribute and Virtual Attribute

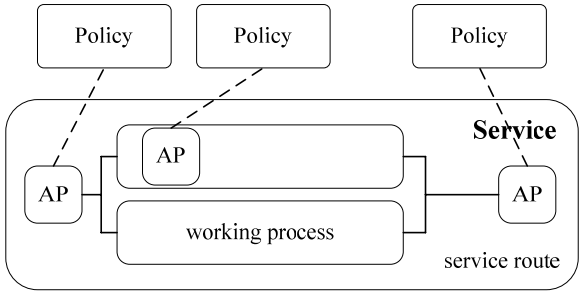


Figure 4. Service control in SODAC model

The virtual attribute is similar to a membership ticket, but different in that membership ticket can be used as identity certificate or authority directly for service invoking, but the virtual attribute only provide an access entrance to acquire authority. So that the attribute is delivered only to whom requires it and when it is needed. This mechanism can prevent malicious user from identity collecting. Once the delegate requester performs the authority delegated, the delegate provider would know and confirms it. Additional policy could also be established for the accessing service to provide constraints on delegation. It will be discussed in the following.

B. Delegation constraint

With virtual attribute, we have restricted the delegation requester from achieving the practical authority himself. In order to place more constraints on the delegation, the delegation provider can issue additional policy on the accessing service. The management to virtual attribute is referring to the template-instance management of task-role-based access control (T-RBAC) [7]. With that, the delegation provider will reserve records of all the delivered virtual attributes. When an attribute request arrives, the delegation provider will first recognize the corresponding virtual attribute that it comes from, and then control the service process according to policy of the original service and that of particular delegation service.

The SODAC model absorbs idea from usage control (UCON) that authentication can be configured to be performed before, simultaneous with or after the service is ongoing [8]. Every policy is attached to an *authentication point* (AP) along the service route, as shown in Figure 4. So that, we can easily define policy to revoke delegation dynamically as soon as its validity expire.

For example, if the delegation provider wants to set a policy to constrain the times that delegated authority can be used. It can be attached to an authentication point set after the service is accomplished. So the policy will be authenticated every time after the service invocation is finished, and once the authentication returns false, the virtual attribute will be revoked.

C. Delegation chain

As discussed in the introduction, the permission to re-delegate authority must be controlled. However, the situation still exists that another party *C* may request relegation from party *B* whose own authority is delegated from party *A*, especially in SOA. With service composition, it is an ordinary form that some service invoking other one in its working process, but is a composing part of another service simultaneously, known as service chain.

Since the actual accessing authority belongs to the original delegation provider, the eventual delegation confirmation should be controlled by it. But the delegation process should be mediated by other delegation party to protect information of the original provider. As the same as many previous research in re-delegation, we introduce in delegation chain to constrain permission of re-delegation.

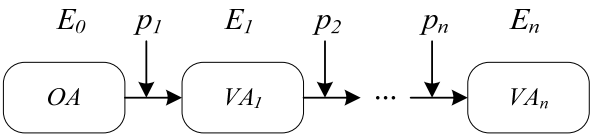


Figure 5. Example of delegation chain

As shown in Figure 5. E_0, E_1,…, E_n are entities; VA_1, VA_2,…, VA_n are virtual attributes; OA is the corresponding original attribute; p_1, p_2,…, p_n are policies. When a new entity request delegation from some delegating party, it will transfer the request along the delegation chain back to original entity. If the original entity approves the request, he will produce a new virtual attribute derived from the virtual attribute be requested, and send it back to delegation requester along delegation chain. The new delegated virtual attribute is restricted by all the policies assigned to its previous ones in delegation chain. So, the longer the delegation chain is, the higher restriction will be set for the new delegation requester. The original delegation provider can also define delegation negotiation policy previously to restrict length or width of the delegation chain. The delegation chain is restored and maintained only at the original delegation provider's side. Once a virtual attribute is revoked, all the other ones derived from it should be revoked too.

D. Delegation process

When the delegation requester is unfamiliar to the delegation provider, a negotiation is necessary to determine whether to approve it or not. The situation is similar to trust negotiation in regular service invoking. The difference is that the final certificate sent back to requester is not membership, but virtual attribute. We introduce in a particular *delegation service* at delegation provider's side. All the services for attributes of the entity can be invoked from the delegation service. So that, different negotiation route can be produced according to the attribute requested. The delegation process is shown in Figure 6.

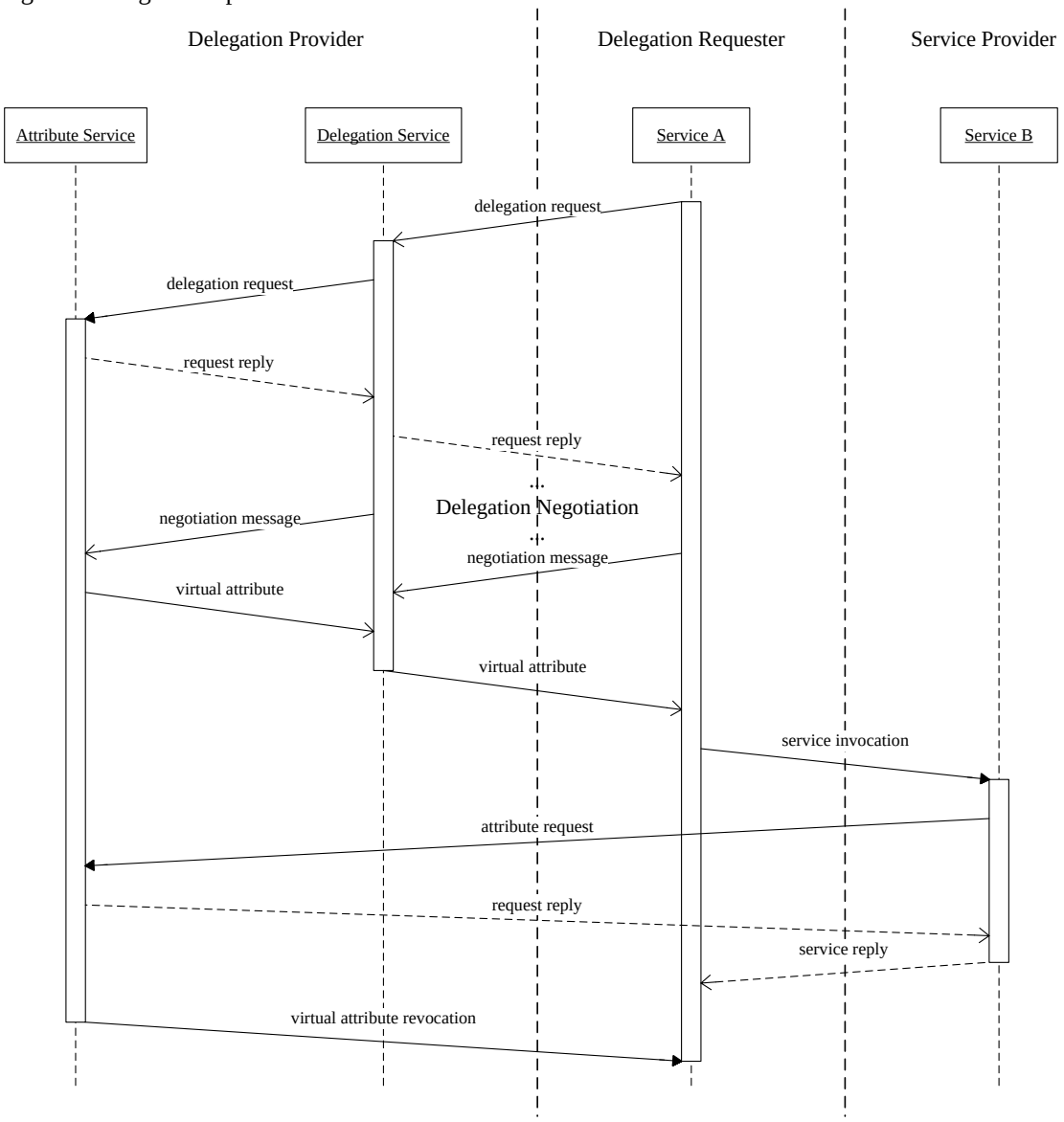


Figure 6. Sequence diagram of delegation process

IV. Conclusion

The NCE is a typical decentralized environment, resources and authorities are distributed all over the involved entities. With SOA, it is easy to share resources with others by issuing services. But the authority to access resources must be controlled to protect sensitive information. In this research, we investigated in the management to delegation at the base of SODAC model, and take the three widely concerned issues in consideration. With the proposed mechanism, we could protect the actual authority from leaking out, control delegation according to the practical requirement, and constrain its validity and re-delegation.

This work is just a stepped accomplishment of our research on SODAC model. We only proposed mechanism to control delegation, practical implementation and validation will be investigated in future work.

References

[1] Wei She, Bhavani Thuraisingham, and I-Ling Yen, "Delegation-based Security Model for Web Services", 10th IEEE High Assurance Systems Engineering Symposium, IEEE CS, 2007, pp. 82-91.

[2] Mitsuhiro Mabuchi, Yasushi Shinjo, Akira Sato, and Kazuhiko Kato, "An Access Control Model for Web-Services That Supports Delegation and Creation of Authority", 7th International Conference on Networking, IEEE CS, 2008, pp. 213-222.

[3] DoD Architecture Framework Version 1.5, Volume I: Definitons and Guidelines. The United States: Department of Defense, 2007, 4.

[4] Han Ruo-Fei, Wang Hou-Xiang, Xiao Qian, Jing Xiao-Pei, and Li Hui, "Service oriented decentralized access control for military systems in Net-Centric Environment", 2009 International Symposium on Electronic Commerce and Security, IEEE, 2009, pp.

[5] Han Ruo-Fei, Wang Hou-Xiang, Li Hui, and Wang Yu-Hua, "Membership-based Access Control for Trust Negotiation in Open Systems", The 5th International Conference on Information Assurance and Security, IEEE, 2009, pp.

[6] R.S.Sandhu, D.F.Ferraiolo, and R.Kuhn, "The NIST Model for Role Based Access Control: Towards a Unified Standard", Proceeding of the 5th ACM Workshop on Role-Based Access Control, ACM press, Berlin, 2000, pp. 47-63.

[7] Sejong Oh, and Seog Park, "Task-role-based access control model", Information Systems, 2003, 28, pp. 533-562.

[8] Jaehong Park, "Usage Control: A Unified Framework for Next Gerneration Access Control", The Gorge Marson University, 2003.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 321-324

Dim Target Detection System Based on DSP

Yongxue Wang, and Jian Zhang
School of Science, Hebei University of Technology, Tianjin, 300401, China
Email: yx_wang@hebut.edu.cn

***Abstract*—The research on distributed system's dim target detection and tracking is very important to improve the performance of infrared guard system. When the image background is very intricate, and there are few pels of targets, it is very hard to check the target. An local texture analysis method which is very fit for dim target detection in low signal to noise rate was adopted to realize the detection. In order to satisfy the real-time realization request of dim target detection, a dim target detection system was presented, and it was based on TMS320DM642 which is a high performance digital multimedia DSP chip produced by TI Company. The program's processing speed and pipelining efficiency was improved by optimizing the software program, so the system has a very good software and hardware system structure. Experiments show that the system can detect dim target with 1-3 pixels in clutter background steadily and in real-time.**

***Index Terms*—dim target, detection, TMS320DM642, hardware system**

I. Introduction

With the developing and applying of digital technology and micro-electronics technology, aviation electronics technology is growing quickly, the fire control system should satisfy the higher request of weapon system of fighter plane. Distributed detection system is a new concept of passive photo-electricity system research in military application, and it's a new developing direction of photo-electricity system in military application. Distributed infrared multi-sensor system uses infrared passive detection and tracking. It can work when radar system is suspended, so it has good ability of anti-electromagnetism and has many characters such as high resolution imaging[1]. At the same time, using several sensors together can highly improve the target detection and tracking performance. The research on distributed system's dim target detection and tracking is very important in improving the performance of infrared guard system and new generation fighter plane's fire control system[2].

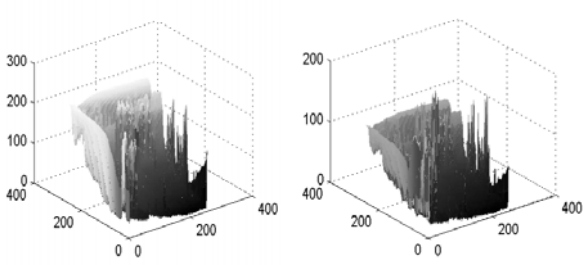
At present, dim targets detection and tracking in clutter infrared image sequences is an important research problem. The algorithm design has reached a relatively mature stage, but there are some bottleneck problems in real-time hardware realization, such as great data quantity and low processing speed, which make it hard to realize in applications. The continuous appearing of high performance DSP chips makes it possible to realize dim target detection and tracking in real-time[3]. We use the high performance DSP chip TMS320DM642 of Texas Instrument(TI) as a core, design and realize an whole real-time hardware system including image input and output, and enhance the processing speed of the system by optimizing software program structure of the system, and the choosing of algorithm becomes very flexible and diversified. The system has a simple structure, a small cubage and a low power consume, so it has very high utility level.

II. A Method of Dim Targets Detection

This kind of new dim targets detection method is based on local texture analysis. Traditional methods always use the gray value difference of target and background in infrared image to realize dim target detection[4]. This method uses the local multi-resolution character vectors difference of target and background to realize detection. It can detect dim target effectively. Experiments on real data prove that it can detect target with 1~3 pixels. It can realize good real-time performance using Mallat algorithm and à trous algorithm in wavelet decomposition phase, and can be realized on hardware.

Figure 1 gives an Original infrared image and its local distance images after these two kinds of wavelet decomposition. The results show that there is a biggish peak value at the part where the local texture vary tempestuously.

(a) an original infrared image



(b) the local distance image after Mallat algorithm

(c) the local distance image afterà trous algorithm

Figure 1. Original infrared image and its local distance images after two kinds of wavelet decomposition

III. CHARACTER OF TMS320DM642

TMS320DM642 is the highest performance digital multimedia processor of TMS320C6X produced by TI, and it is based on C64X core and has Very-Long-Instruction-Word (VLIW) architecture, which is developed by TI[5]. The DM642 offers cost-effective solutions to high-performance DSP programming challenges, with performance of up to 4800 million instructions per second (MIPS) at clock rate of 600MHz. The DM642 uses a two-level cache-based architecture and has a enhanced Direct-Memory-Access (DMA) controller with 64 independent channels. Those are the representation of high-performance, the processing speed can be highly boosted if they can be used and managed reasonly.

DM642 has 548 pins with Ball Grid Array (BGA) package, and it is highly integrated with powerful and diverse set of peripherals: there configurable video ports, providing a glueless interface to common video decoder and encoder devices, and support multiple resolution and video standards; VCXO interpolated control port (VIC); 10/100 Mbs/s ethernet MAC (EMAC); management data input/output (MDIO); multichannel audio serial port (McASP); inter-integrated circuit (I^2C) bus; two multichannel buffered serial ports (McBSP); three 32-bit general purpose timers; configurable 16-bit or 32-bit host port interface (HPI); 32-bit/60MHz 3.3-V peripheral component interconnect (PCI) master/slave interface conforms to PCI specification 2.2; six general purpose I/O pins; 64-bit external memory interface (EMIF), glueless interface to asynchronous memories and synchronous memories.

It is obvious that DM642 is a powerful multimedia processor and can be a good platform of making up of multimedia signal processing: Its diverse set of peripherals make it almost a single chip hardware platform of embed multimedia system; its complete programmable ability makes it compatible with most kinds of multimedia processing standards, and composes a general-use software platform.

IV. HARDWARE SYSTEM DESIGN

A. Hardware System Architecture

The real-time ability of data processing, hardware system dimension and the difficulty of debugging the system all are considered when the system project is designed. The hardware system mainly includes four parts: minimum DSP system, input module, output module and memory module. Figure 2 shows its whole architecture. The analog video signal of infrared image sequence is input into system from infrared sensor in front of the system, digital video signal can be gotten by video decoding in input module, the dim target detection and tracking can be realized on DM642 chip, and the detection and tracking results will be transmitted to the back processor through McBSP serial port in output module, at the same time, results video can be showed on monitor after having been processed by video encoding. The diverse set of peripherals of DM642 are fully used in this system: PCI interface is added into input module, so video images can be directly transmitted into the system through host computer, this can greatly facilitate the algorithm hardware simulation and system debugging at the developing phase of the system.

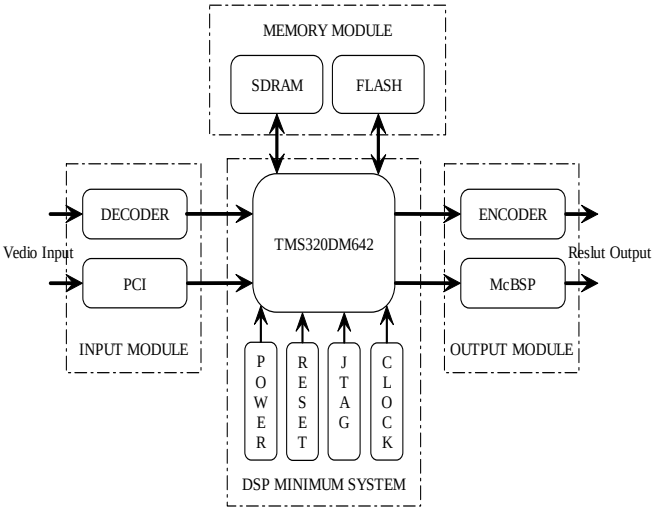


Figure 2. Hardware system architecture

B. Minimum DSP system

The DSP minimum system includes chip DM642 and other circuits which can make it run. Clock circuit uses crystal which can provide 50MHz stable and reliable clock output for DSP input and external memory, the phase lock loop (PLL) is set at ×12 mode (CLKMODE1=1, CLKMODE0=0), the 50MHz external clock input will increase to 600MHz after being processed by PLL; the power circuit uses two chips (TPS54310) to provide 3.3V I/O voltage and 1.4V core voltage, and ensures the power supply timging of the whole system; the reset circuit realizes system reset and supervises each voltage of the system; the JTAG circuit provides interface to external emulator to realize simulation and debugging.

C. Memory module

Memory module includes two parts: SDRAM circuit and FLASH circuit. A great deal of image data should be processed at the phase of dim target detection and tracking, it is requested that there are enough memory space in the system, so two SDRAM chips (MT48LC4M32B2) are parallel connected to EMIF of DM642 in this system, and are mapped at CE0 space, then 64-bit 32M byte data memory and access can be realized, the memory's working clock will be locked at 133MHz, so high speed memory and access of a great deal data can be realized and the real-time ability of the system can be ensured. The FLASH part uses a chip of AM29LV033C, configured as 8-bit width, mapped to CE1 space, and has 4M bit memory capability. This part can be used to load the processing program at the reset moment of the system and realize offline running.

D. Video input module

Video input module includes two alternate schemes: input from infrared sensor or input from host computer by PCI. In practice application, video signal is inputted into system from infrared sensor, after being A/D transferred and video decoded, the digital video is transferred into *DSP* at format *BT656*, chip *SAA7115* is used in this part, it can be connected with the video ports of *DM642* without glue, I^2C Bus is used to configure the video control registers. At algorithm simulation and system debugging phase, digital video data will be transferred into *DSP* from host computer through *PCI*, this interface can realize 32-bit data transfer at speed of 60M*Hz*, which can simulate the analog video input in real-time, so it makes the system more configurable.

E. Result output module

Result output module also includes two parts: result data output and result video show. After the task of dim target detection and tracking has been finished, the result information such as target quantity, location ordinates and target velocity will be gained, then they will be transferred to back processor through the McBSP of DM642 using serial port transfer protocol, at the same time, the result of detection and tracking will be showed on monitor after being transferred to analog video by video encoder chip SAA7105, then we can get intuitionistic result on monitor, so it not only realizes videotext, but also makes it easy to debug the system.

V. Software System Design

A. System work flow

At the process of realizing software scheme design, C language and linear assemble language in CCS (Code Composer Studio)are used to realize image processing program and accomplish dim target detection and tracking, the work flow is showed in Figure 3. After being powered, the system works in the following steps: 1) Program is booted into DSP by FLASH memory, then begin to initiate the system; 2) step into waiting state till a video interrupt occurs, then DSP reads video data; 3) transfer target detection program to detect target, if the target appear, begin to track the target, at the same time, tracking result will be fed back to the detection process to confirm the target; 4) transfer the target quantity, location coordinates and target velocity as detection and tracking result to next processor through serial port, at same time, show result video with target being locked.

B. Detect and track algorithm

Commonly, there are few target pixels in infrared image. And it is always lack of target shape and structure information, at the mean time, there are sensor noise and clutter background in image, target is always totally submerged in the clutter background, so traditional methods can not work very well on detecting and tracking dim target.

A detection method based on local texture analysis is used at detection phases in this system, at first, local texture characters in different resolutions and different directions will be obtained through wavelet decomposition, then a multi-resolution local distance image can be gotten by calculating the distance between every character vector and its local background vector, at last, the potential targets can be detected out through neighbor probability method. After the system has been shifted into tracking phase, a tracking method based on curve fitting prediction is used to track and lock the target. For the targets with velocity less than 1 pixel/frame, the system will segment the image into binary image and lock the targets directly. It makes the system process faster.

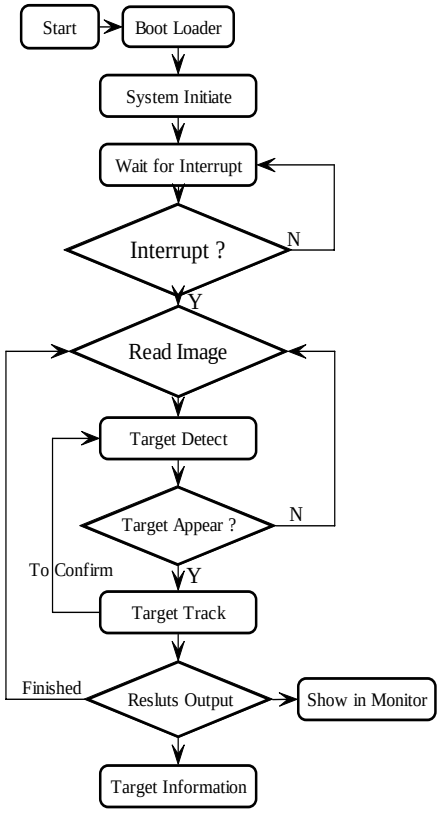


Figure 3. Software system flow

C. Program optimization

In order to enhance the real-time ability of the system, and make it able to fit more complex software algorithms, two aspects of work are finished in this paper: 1) use C language and linear assemble language to realize mixed program, C language is the main frame, linear assemble language appears as inner code or function that can be transferred by C language and it accomplishes the algorithm that consumes most of the CPU clock, which makes the software program able to use DSP resource sufficiently and makes the code more compact and efficient. 2) use enhanced Direct-Memory-Access(DMA) to realize mass image data moving from external memory to inner memory without consuming CPU clock, which greatly lightens the burden of moving mass data of CPU and enhances the efficiency of the system.

VI. Experiment and Conclusion

In order to validate the ability of the dim target detection and tracking system designed in this paper, real image sequence data of America Army Laboratory is used in simulation experiment. The size of every frame in the image sequence is 320×244, and its SNR is about -8db, the image sequence is input into the system from host computer through PCI interface. Dim targets with 1-3 pixels are effectively detected out by detection method based on local texture analysis, on platform of P4 2.8G computer and Matlab6.5, it takes the detection algorithm 0.6410 second to finish detection in every frame. After having migrated the algorithm into the hardware system, the system works stably and realizes detection and tracking effectively. The detect time is reduced to 0.016 second, it's one-forty of primary time, so the system can realize real-time detection and tracking on video with frame speed at 62 frame/second and has very good stability and real-time ability.

The highest performance digital multimedia processor DM642 of TI is used as core processing unit to realize infrared dim target detection and tracking system, with performance of up to 4800 million instructions per second (MIPS) at clock rate of 600MHz. The system has a good software and hardware architecture, which make it easy to satisfy the request of each kind of algorithm. Experiment on real image data shows that the system has very good real-time ability and can realize stability and effectively detect and track of dim target in the video at rate of 62 frame/second in real time. At the same time, the system can be used in hardware simulation and testing of image processing algorithm, and by changing the software program, the system can be widely used in many fields of video image processing, so it has very good re-configurability and versatility.

References

[1] CUI Yu-ping, ZHENG Sheng, LIU Yong-cai. SVM-based infrared small target detection[J]. Infrared and Laser Engineering, 2005, 34(6):696-702

[2] WANG Yue-huan, CHENG Sheng-lian, ZHOU Xiao-wei, et al. Multi-scale small targets detection in clutter based on multi-level filter[J]. Infrared and Laser Engineering, 2006, 35(3): 362-366

[3] XU Bin, ZHENG Lian, WANG Keyong, et al. Dim targets detection method based on local texture[J]. Opto-Electronic Engineering, 2005.32(1):40-42

[4] MALLAT S. A Wavelet Tour of Signal Processing [M]. Second Edition. Beijing: Academic Press, 1999: 115-117

[5] TMS320DM642 Video/Imaging Fixed-Point Digital Signal Processor Data Manual[Z]. Literature Number: SPRS200E. Texas Instruments Incorporated (TI). July 2002-Revised March 2004.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 325-329

Research of Routing System which applied in ASP.NET MVC Application

Xiangjun Li [1, 2], Liang Huang [1], Zhenrong Lin [1], and ZiLong Sai [1]
[1] Computer Department, College of Information Engineering, Nanchang University, Nanchang, China
Email: brilliance_h@live.cn
[2] Colleges of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
Email: lixiangjun@chinalmtc.com

***Abstract*—Most URLs have been correspond directly to files on the server side in traditional ASP.NET WebForms and many other web platforms. Though, it's very easy to access the files that existed on the server, through inputs its name directly in the URL, which is easy to understand, but it also makes the files in insecurity, This paper has analyzed the control from Routing System to URL schemas, which constructs agility, concise and beautiful URLs, Then apply it to an Online Store System that designed in ASP.NET MVC application as a demonstration which aims to make search engine optimization.**

***Index Terms*—ASP.NET MVC, Routing System, Routing mechanism, URL Schema**

I. INTRODUCTION

Routing System is all about the Universal Resource Locator (URL) and how it is used as an external input to the applications. The URL has led a short but troubled life and the HTTP URL is currently being tragically misused by current web technologies. As the web began to change from being a collection of hyperlinked static documents into dynamically created pages and applications, the URL has been kidnapped by web technologies and it often emerges undesirable mistakes. The URL is in trouble and as the web becomes more dynamic, software developers have a chance to introduce a rescue operation and bring back the simple, logical, readable and flexibility URL.

Routing System brings the chance to software developers to take URL back and have controlling over the URLs of applications. The core conception of routing in traditional ASP.NET WebForms and many other web platforms has been that URLs correspond directly to files on the server. The server implements and serves the page or file corresponding to the incoming URL, while in ASP.NET MVC; URLs are not expected to correspond to files on the web server. In fact, that wouldn't even make sense—since ASP.NET MVC's requests are handled by controller classes (compiled into a .NET assembly), there are no particular files corresponding to incoming URLs [1].

While routing is not core to all implementations of the MVC pattern but is often found implements as a convenient extra way to add an extra level of separation between external inputs and the controllers and actions which make up an application.

II. BASIC THEORY OF ROUTING SYSTEM

In ASP.NET MVC application, software developers are given complete controlling of their URL schema—that is, the set of URLs that are accepted, and whose mappings to controllers and actions. This schema isn't restricted to any predefined pattern and doesn't need to contain any file name extensions or the names of any of the classes or code files.

A. Routing mechanism

The routing mechanism runs early in the framework's request processing pipeline. Its job is to take an incoming URL and use it to obtain an IHttpHandler object that will handle the request.

A routing system in ASP.NET MVC framework is responsible for managing the decoupling of the URL with the application logic [2、3]. It must manage this in both directions so that it can:

- Map URLs to a controller/action and any additional parameters.
- Construct URLs which match the URL schema from a controller, action and additional parameters.

This is more commonly referred to as inbound routing (Figure 1) and outbound routing (Figure 2), respectively. Inbound routing describes when a URL ends up invoking a controller action, whereas outbound routing describes the framework generating URLs for links and other elements on the web site.

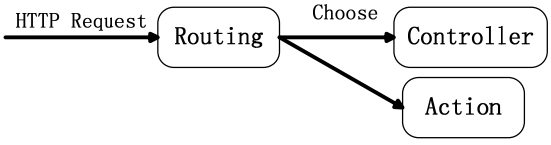


Figure 1. Inbound routing takes an HTTP Request (a URL) and maps it to a controller and action

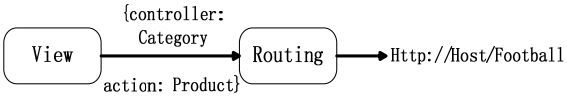


Figure 2. Outbound routing generate appropriate URLs from a given set of route data (controller & action)

When the routing system performs both of these tasks, the URL schema can truly be independent of the application logic and as long as it is never bypassed when

constructing links in a view then the URL schema should be trivial to change independently of the application logic.

B. Three main characters of routing

Routing configurations are built up of three main elements: they are RouteBase, Route and RouteCollection [4-6].

- RouteBase is the abstract base class for a routing entry. Software developers can implement unusual routing behaviors by deriving a custom type from it.
- Route is the standard, commonly used subclass of RouteBase that brings in the notions of URL template, defaults, and constraints.
- A RouteCollection is a complete routing configuration. It's an ordered list of RouteBase derived objects (e.g. Route objects).

RouteTable.Routes is a special static instance of RouteCollection. It represents the application's actual, live routing configuration. Typically, it populated just once, when the application first starts, during the Application_Start() method in Global.asax.cs. It's a static object, so it remains live throughout the application's lifetime, and is not recreated at the start of each request [7].

Normally, the configuration code isn't actually inline in Application_Start(), but is factored out into a public static method called RegisterRoutes(). That makes the configuration easy to access from the automated tests.

To see how routing system works clearly. It creates a blank new ASP.NET MVC project, which executing codes are as the following:

```
Route myRoute = new Route
("{controller}/{action}/{id}", new MvcRouteHandler()){
    Defaults = new RouteValueDictionary (new {
    controller = "Home", action = "Index", id = ""})
    };
routes. Add ("Default", myRoute);
```

Now, the software developers are given complete controlling of their URL schema—that is, the set of URLs that are accepted, and their mappings to controllers and actions. This schema isn't restricted to any predefined pattern and doesn't need to contain any file name extensions or the names of any of classes or code files. Table I shows an example:

TABLE I. HOW THE DEFAULT ROUTE ENTRY MAPS INCOMING URLS[8]

Incoming URL	Might Correspond To
/	{controller = "Home", action = "Index", id = " "}
/Category	{controller = "Category", action = "Index", id = " "}
/ Category /Basketball	{controller = "Category", action = "Basketball", id = " "}
/ Category / Basketball /16	{controller = "Category", action = "Basketball", id = "16"}

C. Routing configuration

There are five properties software developers can configure on each RouteTable entry. These affect whether or not it matches a given URL, and if it does, what happens to the request, see Table II.

TABLE II. PROPERTIES OF SYSTEM.WEB.ROUTING.ROUTE[9]

Property	Meaning	Type	Example
Url	The URL to be matched, with any parameters in curly braces (required).	string	"Browse/{category}/{pageIndex}"
RouteHandler	The handler used to process the request (required).	IRouteHandler	new MvcRouteHandler()
Defaults	Makes some parameters optional, giving their default values.	RouteValueDictionary	new RouteValueDictionary(new { controller = "Products", action = "List", category = "Fish", pageIndex = 3 })
Constraints	A set of rules that request parameters must satisfy.	RouteValueDictionary	new RouteValueDictionary(new { pageIndex = @"\d{0,8}" })
DataTokens	A set of arbitrary extra configuration options that will be passed to the route handler (usually not required).	RouteValueDictionary	

D. The order of Route Entries

RouteCollection is an ordered list, and the order in which we add route entries is critical to the route-matching process. The system does not attempt to find the "most specific" match for an incoming URL; its algorithm is to start at the top of the route table, check each entry in turn, and stop when it finds the first match [10]. So put more specific route entries before less specific ones which configure as follows:

```
routes.MapRoute (
    "Specials",   // Route name
    "DailySpecials/{date}",          // URL with
    parameters
    new {controller = "Catalog", action =
    "ShowSpecials" });    // Parameter defaults
routes.MapRoute (
    "Default",    // Route name
    "{controller}/{action}/{id}",    // URL with
    parameters
    new { controller = "Home", action = "Index", id =
    "" });          // Parameter defaults
```

E. Unit Testing Routes

One of the core design principles of the ASP.NET MVC Framework is enabling great testing support. Ref. [11] like the rest of the MVC framework, programmers can easily unit test routes and route matching rules. The MVC Routing system can be instantiated and run independent of ASP.NET-which means programmers can load and unit test route patterns within any unit test library and using any unit test framework (NUnit, MBUnit, MSTest, etc)[12].

Although programmers can unit test an ASP.NET MVC Application's global RouteTable mapping collection directly within their unit tests, in general it is usually a bad idea to have unit tests ever change or rely on global state [13]. A better pattern that programmers can use is to structure their route registration logic into a RegisterRoutes () helper method like below that works against a RouteCollection that is passed in as an argument (the unit tests code below is compiled based on the mock):

```
private string GetOutboundUrl(object routeValues){
   //Get route configuration and mock request context
   RouteCollection routes = new RouteCollection ();
   MvcApplication.RegisterRoutes (routes);
   var mockHttpContext = new Moq.Mock
<HttpContextBase>();
   var mockRequest = new Moq.Mock
<HttpRequestBase>();
   var fakeResponse = new FakeResponse();
   mockHttpContext.Setup(x=>x.Request).Returns(mock
Request.Object);
   mockHttpContext.Setup(x=>x.Response).Returns(fake
Response);
   mockRequest.Setup(x=>x.ApplicationPath).Returns("/
");
   //Generate the outbound URL
   var ctx=new RequestContext(mockHttpContext.Object,
new RouteData());
   return routes.GetVirtualPath(ctx, new
RouteValueDictionary(routeValues)).VirtualPath;}
```

Then programmers can write unit tests that create their own RouteCollection instance and call the Application's RegisterRoutes () helper to register the application's route rules within it [14-16]. They can then simulate requests to the application and verify that the correct controller and actions are registered for them without having to worry about any side-effects.

III. EXAMPLES OF APPLICATION

Now, it's been aware of the principle of the Routing mechanism and the theory behind its configuration. It's necessary to put the mechanism into action for demonstration and see how those benefits work out in a demonstration application.

The demonstration application, SportsStore, is designed by ASP.NET MVC Framework for online shopping. There is a product catalog browsable by category and page index between which would be used by routing system to navigation or redirection from one product to another.

A. Add RouteTable entries

The default route (matching{controller}/ {action}/{id}) is so general in purpose that we could build an entire application around it without needing any other routing configuration entry[17、18]. However, if it does want to handle URLs that don't bear any resemblance to the names of controllers or actions, then it will need other configuration entries, these route entries show in follows (Just show representative ones):

```
routes.MapRoute (
      null,
      "{category}",          //Matches ~/Football or
      ~/Anything with no slash
      new { controller = "Products", action = "List", page
      = 1 });
routes.MapRoute (
      null,
      "{Category}/Page {page}",//Matches
      ~/Football/Page567
      new { controller = "Products", action = "List" },
          //Defaults
      new { page = @"\d+" });       //Constraints: Page
      must be numerical
```

This entry will match /Catalog or /Catalog? some=querystring, but not /Catalog/Anythingelse. It's understood which parts of a URL are significant to a Route entry.

B. URL Patterns Match the Path Portion of a URL

When a Route object decides whether it matches a certain incoming URL, it only considers the path portion of that incoming URL. That means it doesn't consider the domain name (also called host) or any query string values. Figure 3 depicts the path portion of a URL [19、20].

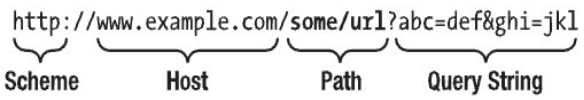


Figure 3. Identifying the path portion of a URL

The URL pattern "Catalog" would therefore match both http://a.b.c.d:1234/Catalog?query=string and http://example.com/Catalog. If programmers deploy to a virtual directory, their URL patterns are understood to be relative to that virtual directory root.

C. Order Route Entries

Order all of the RouteTable Entries according to the golden rule of routing--put more specific route entries before less specific ones [21]:

```
new { controller = "Products", action = "List", category
= (string)null, page = 1 },
new { controller = "Products", action = "List", category
= (string)null },
new { controller = "Products", action = "List", page =
1 },
new { controller = "Products", action="List" }
null,
```

D. Demonstrate

After adding route entries, they are ordered and configured in ASP.NET MVC application, it has broken the traditional URL schemas, convert the common URL to be human-friendly URL, See Figure 4:

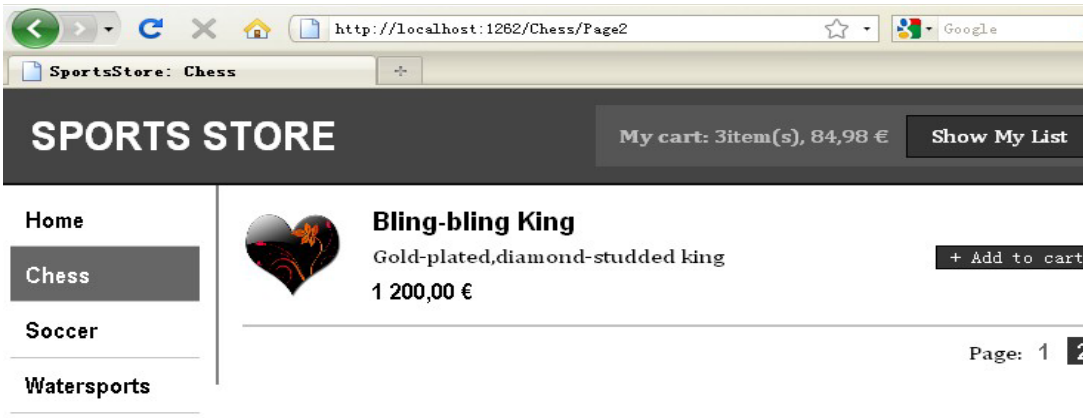


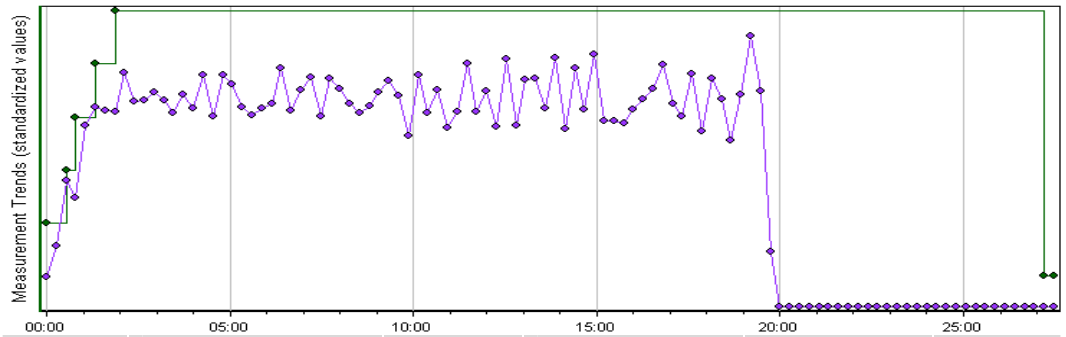
Figure 4. Clear URL designed by routing system in ASP.NET MVC Application

It shows that the URL address is at the Page2 with clear, human-friendly and simple, it presents the URL with the correspond information about the product, which let users understand immediately, when redirected to a product detail page.

E. Comparison

In Figure 5, we can see the common URL which contains extension name of asp file in the ASP.NET WebForm application.

Figure 5. URL in the traditional ASP.NET WebForm



Color	Graph	Scale	Measurement	Correlation Match	Average
	Route System of MVC	Standardized	Reaction、Interaction	100	N/A
	ASP.NET WebForm	Standardized	Reaction、Interaction	76	103.756

Figure 6. Contrast Results from the Route System of MVC Application and the traditional ASP.NET WebForm by LoadRunner Analysis Software

Compared to ASP.NET WebForm, the routing system of MVC application is better than ASP.NET's at the following:

- It's doesn't use the storage path on the server as the URL, MVC routing system configures the URL by controller.
- It hides the complex .aspx extensions which make sure the files on the server's security.
- It could be designed a completely independent web application and attached to the core routing system with its placeholder, defaults, route validation, and URL generation features.

Through LoadRunner analysis software, it can analyze the ASP.NET MVC application and the traditional ASP.NET WebForm, Compared to them both in the reaction time and interaction by users. It shows that the routing system of MVC application is better than the ASP.NET's, See Figure 6.

IV. CONCLUSION

To design the URL schema for an application is a great challenge which covered in this paper. There is never a definitive answer to that route will need to be implemented and that the code needed to generate routes and URLs from routes is very simple, while the process of designing of schema is not. Ultimately every application is different, some will be effective that the default routes created by the project template where others will have route manager classes stretching to hundreds of lines.

All descriptions above shows that order of definition matters and that careful consideration have to be made when adding new routes to the application. As more routes are defined, the risk of breaking existing URLs increases.

Separation of the URL schema from the underlying code architecture gives ultimate flexibility and allows software developers to focus on what would make sense for the users on the URL rather than what the layout of our source code requires.

ACKNOWLEDGMENT

This paper is supported by the 2009 Jiangxi Natural Science Foundation (No.2009JX02289) and by the 2008 Technical Plan Project of Jiangxi Education Department (No.GJJ08036).

REFERENCES

[1] Steven Sanderson, Pro ASP.NET MVC Framework, 2009, pp.221-258.

[2] Jeffrey Palermo, Ben Scheirman, Jimmy Bogard, ASP.NET MVC in Action, 2007, pp.81-100.

[3] Rob Conery, Scott Hanselman, Phil Haack, Scott Guthrie. 2009. Professional ASP.NET MVC 1.0.199-210.

[4] Dino Esoisito, Microsoft® ASP.NET and AJAX: Architecting Web Applications, 2009, pp.70-73.

[5] Steven Sanderson, Pro ASP.NET MVC V2 Framework, unpublished.

[6] Jon Flanders, RESTful .NET: Build and Consume RESTful Web Services with .NET 3.5,2008, pp.219-221.

[7] Stephen Walther, ASP.NET MVC Framework Unleashed, 2009, pp.275-284.

[8] Simone Chiaretta, Keyvan Nayyeri, Beginning ASP.NET MVC 1.0, 2009, pp.131-140.

[9] Nick Berardi, Al Katawazi, Marco Belinaso, ASP.NET MVC 1.0 Website rogramming: Problem-Design-Solution, 2009, pp.30.

[10] Enad Ibrahim, ASP.NET MVC 1.0 Test Driven Development: Problem-Design-Solution, 2009, pp. 73-92.

[11] Roy Osherove, The Art of Unit Testing: With Examples in .Net, 2009, pp.132-135.

[12] Scott Guthrie, ASP.NET MVC Framework (Part 2): URL Routing, http://weblogs.asp.net/scottgu/archive/2007/12/03/asp-net-mvc-framework-part-2-url-routing.aspx, 2007.

[13] Jeff McWherter, Ben Hall, Testing ASP.NET Web Applications, 2009, pp.64-71.

[14] Imar Spaanjaars, Beginning ASP.NET 3.5: In C# and VB, 2008, pp.520-525.

[15] Karli Watson, Christian Nagel, et al. Beginning Microsoft Visual C# 2008, 2008, pp.326-330.

[16] Maarten Balliauw, ASP.NET MVC 1.0 Quickly, 2009, pp.79-90.

[17] Matthew MacDonald, Mario Szpuszta, Pro ASP.NET 3.5 in C# 2008, 2008, pp.166-177.

[18] Andrew Troee. 2008. Pro C# 2008 and the .NET 3.5 Platform, Fourth Edition, 2008, pp.309-332.

[19] Microsoft ASP.NET MVC Tutorials. ASP.NET MVC Routing Overview (C#). http://www.asp.net/(S(ywiyuluxr3qb2dfva1z5lgeg))/learn/mvc/tutorial-05-cs.aspx.

[20] Alex Horovitz, Chris Sutton, et al. Programming ASP.Net MVC, unpublished.

[21] Dino Esoisito, Andrea Saltarello, Microsoft® .NET: Architecting Applications for the Enterprise, 2008, pp.281-285.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 330-334

Building a Speaker Recognition System with one Sample

Mansour Alsulaiman, Ghulam Muhammad, Yousef Alotaibi, Awais Mahmood, and Mohamed Abdelkader Bencherif
Computer Engineering Dept., College of Computer and Information Sciences
King Saud University, Saudi Arabia
msulaiman@ksu.edu.sa; ghulam@ccis.ksu.edu.sa; yaalotaibi@ksu.edu.sa ; awais.mahmood@gmail.com; mbencherif1@yahoo.com

***Abstract*— Speaker recognition system is the process of automatically recognizing the person from his/her speech. To correctly recognize a speaker by the system, many speech samples are needed at different times from each speaker. However, in some applications, such as forensic, the number of samples for each speaker is very limited. In this paper, a method is proposed to train the speaker recognition system based on only one speech sample. From that one sample, other samples are generated. The intent is to provide a complete speaker recognition system, without bothering the speaker to record the speech samples at different times. For this purpose, the speech samples are modified without altering the pitch and the speaker dependent features. Many techniques are used to generate new samples and apply these to the system, when the recognition system is based on the hidden Markov model. The system is built using the HTK software which is a hidden Markov model kit, and the best recognition rate is 85.86%.**

***Index Terms*— Speaker recognition; sample generation; hidden Markov model.**

I. INTRODUCTION

Biometric systems are roughly divided into behavioral and physical pattern measurements. Many countries are making valuable scientific reports on the feasible and viable methods to be used in access or recognition systems. The complexity comes from many issues, the most important ones concern: the no error reproducibility of the registered pattern; the lower data collection error rate, the high user acceptability [1,2], the size of the database [9], the necessary technology to embed into the terminal capture points. These major strategic points tend to classify the different biometric issues into classes, and weight some techniques among others. From the dynamic methods, considered sometimes as changing over time, speech is extremely concerned, as fingerprints are, by the data collection error, many sessions must be executed in order to get a candidate set of samples.

The beauty of speech is its non-invasive nature, i.e. it can be recorded without the person's acceptance, or sometimes without his/her physical attendance. This is also subject to the existing speech recording technology, by the use of sophisticated microphones, or via channels like the landline or mobile phones, or from some TV interviews or radio broadcasts. Unfortunately, sometimes there is not enough speech data recorded, which leads to a lack of enough training data for the model to be correctly trained, that results into a very low recognition rate.

Some methods of speech lengthening are used in human speech recognition for the benefit of speech perception, and a source of information in understanding prosodic organization of speech [3], and also for children in kinder gardens for word discrimination [4].

Regarding the above mentioned problem, different techniques have been proposed in this paper for generating more new samples by using one sample. One method is an expansion or a meaningful lengthening, by modifying one of the existing samples, in order to strengthen the template establishment during training. All other original samples will be used for testing the system.

The paper is organized as follows: section II describes the database, and selection of data; section III defines the modeling technique used in this paper; section IV introduces the front-end processing part of the system, section V illustrates different generation methods which will be explored in this paper, section VI describes the experiments performed with the results given in section VII; in section VIII, the results are analyzed. Finally, section IX concludes the paper and gives suggestions for future work.

II. DATABASE

This research has been conducted with a local dataset recorded at King Saud University, College of Computer and Information Sciences-CCIS, during the year 2007 [5]. The dataset consists of 91 speakers, pronouncing the word "نعم", which stands in English for the word "yes", in 5 different occurrences.

The speaker recognition system is phoneme based, and uses the phonemes of the word "نعم", for recognizing the speaker. The main characteristics of this word are of two aspects. The first aspect is that approximately all the Arab speakers frequently say "yes" (in Arabic) in any discussion. The second aspect is the richness of this word in the phonetic structure. It contains the nasal phoneme [ن], a very pertinent phoneme [ع], at last a bilabial phoneme [م], allowing the capture of the energy of the whole word. It also contains two occurrences of the vowel (فتحة). This richness, plus the fact that it is a commonly pronounced makes it a good choice for our investigation.

The samples will be denoted as:

- First original sample: O_1.

- Four other original samples are used for testing: O_2, O_3, O_4, O_5.

In this work, a part of the database is used. This part consists of 25 different male speakers (20 adults + 5 children). All are native Arabic speakers. Each Speaker uttered the same isolated word نعم five times. The speakers recorded their speech samples in one or two sessions.

III. MODELING TECHNIQUE AND SPEECH FEATURES

In text dependent applications, where there is a strong prior knowledge of the spoken text, additional temporal knowledge can be incorporated by using the Hidden Markov Models (HMMs). HMM is a stochastic modeling approach used for speech/speaker recognition. It is similar to a finite state machine. Each state (node) has an associated probability density function (PDF) for the feature vectors. Moving from one state to another is subject to a transition probability. The first and the last states are not emitting states, since the first state is always from where the state machine starts and the last state is the one, where it always ends, i.e. there are no incoming transitions into the start state and there are no output transitions from the end state. Every emitting state has a set of outgoing transitions and the sum of the probabilities for those transitions is equal to one, since the transition from a non-final state must always occur [6, 7, 8].

The HMM system is build using the HTK (Hidden Markov Toolkit) software, which was developed by Steve Young at Cambridge University in 1989.

Our work involves with three active states, left to right. Also, each state has one mixture. Each phoneme in the keywords is modeled by one model with number of speakers. For example, for a given speaker, each phoneme is modeled differently, even by dealing with the same linguistic sound. These models can be used to find the speaker identity. The silence model is also included in the model set. In a later step, the short pause is created from and tied to the silence model. This system is similar to our original work as presented in [5].

IV. FRONT-END PROCESSING

This step deals with the extraction of features, where speech is reduced into a smaller amount of important characteristics, represented by a set of vectors, such as the Mel Frequency Cepstral Coefficients (MFCC). The cepstral features are mostly used in speaker recognition, due to many reasons, their robustness to noise distortion, their capability to filter the sound as does the human cochlear system, and their degree of de-correlation. The parameters of the system are 16KHz sampling rate with a 16 bit sample resolution, 25 milliseconds Hamming window duration with a step size of 10 milliseconds, and 12 MFCC coefficients as features

V. PROPOSED METHODS

The methods or techniques to generate new speech samples from one original speech sample are proposed. These new samples can be used for the training of speaker recognition systems without altering the speaker identity, such as modifying the pitch and/or the speaker dependent features. All the samples are generated by modifying the first speech sample of each speaker, in the time domain using the PRAAT software. The new samples are generated by any/or combination of the following methods:

A. Copying a part of speech & concatenating it

The samples are generated by copying a small part from the initial speech sample and then inserting it just after the selection. This is done on the first, middle, and last parts of the sample, resulting in three different additional samples. The cut part is around 20 to 30 milliseconds for first group of three new samples, and 40 to 60 milliseconds for the second group.

B. Reversing of the word

In this category four different kinds of samples are generated. The first sample is generated by reversing the original sample. The second, third and fourth samples are generated by coping a small part (approximately 20ms to 30 ms) from the phonemes "م" , "ع" and "ن" then inserting it just after the selection in the reversed word.

C. Adding noise at different SNR

A total of six samples are generated. First three samples are generated by adding babble noise of 5db, 10db and 20db SNR, respectively. The other three samples are generated by adding train noise of 5db, 10db and 20db SNR, respectively.

VI. EXPERIMENTS

In order to confirm that the new generated samples contain supplementary information about the speakers, initially a test experiment is performed on which a system is trained and tested with the same original sample. This

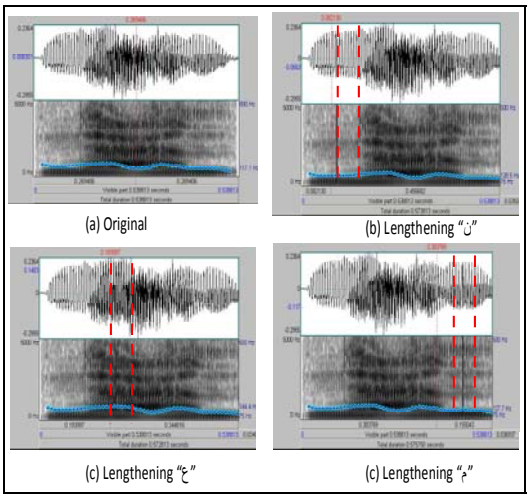


Figure1. Original and generated samples using concatenation.

experiment is named as E_1. The recognition rate is 10%, as expected, which is very low. This is due to the fact, that there is not enough information in one sample. Moreover, the system is trained and tested with the original and generated samples (experiments E_2 and E_3 from section V. A), and 100% accuracy is obtained. This high recognition rate is due to supplementary or additional information obtained during the training by the new generated samples.

However, this is not a real test, because the system shall be tested with other original samples, as with the following experiments.

A. Concatenation

Samples S_5, S_6 and S_7 are generated in this section. These are generated by copying the central part, approximately 20 ms to 30 ms, of each phoneme "ن", "ع" and "م" of the original sample (O_1), then inserting it just after the selected part, respectively. The vertical dotted lines in theFig. 1 show the inserted part. This group is named as conc1.

The samples S_8, S_9 and S_{10} are generated by copying a part of 40 ms to 60 ms. It is a longer length than conc1, of each phoneme "ن", "ع" and "م" of the original sample O_1. Then it is inserted just after the selected part. This group is named as conc2, as mentioned in Table1.

B. Generating Samples by Reversing

Different samples are generated in this part. The first sample in this group S_{11} is generated by reversing the sample O_1. The second, third and fourth generated samples in this group are S_{12}, S_{13} and S_{14}. These are generated by copying a part of approximately 20 ms to 30 ms of each phoneme "م", "ع" and "ن", of the sample S_{11}, then inserting it just after the selected part respectively.

Note that the order of the phonemes is reversed leading to a new word meaning "all together". This group is named as rev4. S_{11} will name as rev1.

C. Generating samples by adding noise

A total of six samples are generated in this last category. The samples S_{15}, S_{16} and S_{17} are generated by adding the babble noise at 5db, 10db and 20db SNR respectively. This group's name is nois1. The three other samples S_{18}, S_{19} and S_{20} are generated by adding the train noise at 5db, 10db and 20db SNR respectively. nois2 is the selected name for this group.

VII. RESULTS

Table 2 describes the results of the different conducted experiments. These experiments are performed using 25 speakers of the database (Section II). In all these experiments the training samples are O_1, and the groups of generated samples as presented in Table 1.

A. Effect of concatenation

Three experiments are conducted in this part:

1. E_4 represents the training of the system by using the samples of conc1; the recognition rate is 50%.

TABLE 1. Techniques for generating samples

Sample Code	Category	Method to generate the new sample	
O_1		Original Sample	
S_5,S_6,S_7	conc1	A small part of the first second and third phoneme which are "ن", "ع" and "م", (approx. 0.02 to 0.03 seconds) is copied and inserted it just after the selection.	
S_8,S_9,S_{10}	conc2	A small part of the first second and third phonemes which are "ن", "ع" and "م", (approx. 0.05 to 0.06 seconds) is copied and inserted it just after the selection.	
S_{11}	rev4	rev1	Reverse of O_1
S_{12},S_{13}, S_{14}			A small part of S_{11}, the first second and third phoneme which are"م", "ع" and "ن", (approx. 0.02 to 0.03 seconds) is copied and inserted it just after the selection respectively
S_{15},S_{16}, S_{17}	nois1	Babble noise is added at 5db, 10db and 20db in the original speech signal O_1.	
S_{18},S_{19}, S_{20}	nois2	Train noise is added at 5db, 10db and 20db in the original speech signalO_1.	
O_2,O_3,O_4,O_5		Original samples	
O_6,O_7		Reverse of O_2,O_3 respectively	

Table 2. Experimental Results

Exp no.	Technique	Training Samples	Test Samples	Rec. rate
E_4	conc1	O_1,S_5,S_6,S_7	O_2,O_3	50%
E_5	conc2	O_1, S_8,S_9, S_{10}	O_2,O_3	40%
E_6	conc1,conc2	O_1, S_5, S_6, S_7, S_8, S_9,S_{10}	O_2,O_3, O_4,O_5	83%
E_7	conc1,nois1	O_1 S_5, S_6, S_7, S_8, S_{15},S_{16},S_{17}	O_2,O_3 ,O_4,O_5	82.11%
E_8	nois1,nois2, conc1	O_1,S_{15}, S_{16}, S_{17}, S_{18} , S_{19},S_{20},	O_2,O_3, O_4,O_5	82%
E_9	conc1,rev4	O_1, S_5, S_6, S_7 , ,S_{11} , S_{12},S_{13},S_{14}	O_1,O_2, O_6,O_7	74%
E_{10}	conc1,conc2, rev4	O_1,S_5,S_6,S_7,S_8, S_9, S_{10},S_{11},S_{12}, S_{13}, S_{14}	O_2,O_3, O_4,O_5	76.77%
E_{11}	conc1,conc2, rev1	O_1,S_5,S_6,S_7,S_8,S_9,S_{10},S_{11}	O_2,O_3, O_4,O_5	85.86%

2. E_5 describes the training of the system by using the samples of conc2; the recognition rate is 40%; this reduced the previous recognition rate by 10 %.
3. In the experiment E_6, it is observed that when both types of concatenations (more information) are included, the recognition rate increased to 82%.

These results indicate that different types of concatenation or more samples will improve the recognition rates, better than a single type of concatenation.

B. *Effect of Noise*

Two experiments are conducted in this part:

1. E_7 illustrates the training of the system by using the samples of conc1 and nois1, the recognition rate is 82.11%.
2. E_8 represents the training of the system by using the samples of nois1 and nois2, the recognition rate is 82%.

C. *Effect of Reverse*

In this part, the used samples are generated as described in section V. B. In the following three experiments

1. E_{10} shows that by training the system using the samples of conc1, conc2 and rev4. The recognition rate is 76.77%.
2. In E_{11}, the samples conc1, conc2, and rev1 are used, the recognition rate increased to 85.86 %.

VIII. DISCUSSION

Experiment E_1 sets the baseline for this work, since it is shown that without enough information in different samples, the HMM will not be able to build a model and recognize it. Repeating the same sample does not give any new information. Then, by conducting experiments E_4 and E_5, it is proved that by careful modification of a sample, new samples can be generated this would give to the HMM more information, and allows building an improved model that enhances the recognition rates. So, these three experiments (E_1, E_4, *and* E_5) are the bootstrap of our work.

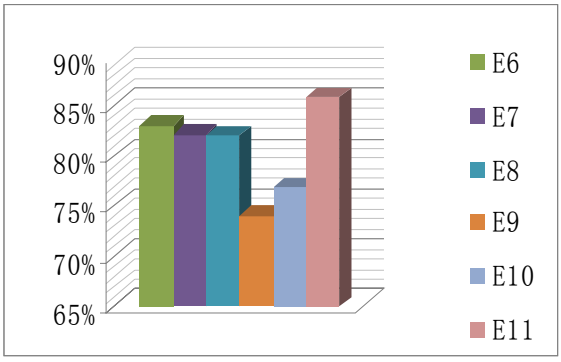


Figure2. Recognition rates per experiment

From experiment E_6, it can be seen that by complementing one method of generation with another the recognition rate increased from 40-50% to 83%. Similar conclusion can be obtained from E_7, where we complemented concatenation with adding noise. Experiment E_8 not only emphasizes this conclusion, but it is also a major result, since it gives a high recognition rate with the samples generated by adding noise and with a little alteration in the original sample (conc1).

These results wrap up with the following consequences, lengthening the vowel part duration may increase the recognition rate. Adding noise together with lengthening vowels may reduce the error rate. E_9 and E_{10} do not give good results as in E_6-E_8. This can be attributed to the fact that there is a co-articulation effect, and the phonemes are context dependant, these are affected by the previous and following phonemes.

In E_{11}, conc1, conc2 and rev1 are used; this experiment outperforms the other methods. The recognition rate attained is 85.86%. Although conc1 and conc2 are generated using the same technique, but they are of different lengths, giving different information. rev1 is generated by reversing the word. Hence, there are 3 methods, which are complementing each other, and results into the best recognition rate. In other words reversing the sample in time domain may have a positive effect on the recognition rate.

This concept of complementary methods give better results, it can be clarified by the following example and analogy. By looking at a view from different angles, we can produce a better picture of the view or even a complete 360 degrees picture. Similarly using different methods of generating new samples will give HMM a better representation, so a better model is built. This point suggests that we investigate other ways of speech samples generation and explore different combinations.

IX. CONCLUSION

Different techniques to generate new samples from an original sample, to overcome the problem of a limited database are proposed. Experimental results showed that even adding different types of noises at different SNR levels to the original sample, during training significantly improved the recognition accuracy.

The experiments also demonstrate that by using different methods to complement each other will lead to an increase in the recognition rate. The highest obtained recognition rate is 85.86% by using samples from three different methods. In the future, we will investigate these techniques on a large number of speakers, as well as improve the accuracy using other possible techniques. Initial results with 50 speakers are encouraging.

The work could be extended to look for the minimum number of words, with some specific phonemes, that may keep the recognition rate as high as possible. This might be used to select a very accurate set of words (phonemes) that characterizes the speaker, without making long sessions of recording. From these selected set of basic words (phonemes) one can generate new samples, using the methods presented in this paper.

REFERENCES

[1]. J. Wayman, A. Janil, D. Maltoni, D. Maio, "Biometric Systems, Technology, Design, and performance evaluation", Springer Editions.

[2]. Thomas Ruggles, "comparison of biometric techniques", http://www.bioconsulting.com/bio.htm.

[3]. Cao Jianfen, "Restudy of segmental lengthening in Mandarin Chinese", ISCA2004.

[4]. Segers, Eliane; Verhoeven, Ludo, "Effects of Lengthening the Speech Signal on Auditory Word Discrimination in

Kindergartners with SLI", Journal of Communication Disorders, v38 n6 p499-514 Nov-Dec 2005.

[5]. S.S. Al-Dahri, Y.H. Al-Jassar,Y.A. Alotaibi, M.M. Alsulaiman, K.A.B Abdullah-Al-Mamun, "A Word-Dependent Automatic Arabic Speaker Identification System" Signal Processing and Information Technology, ISSPIT 2008. IEEE, pp. 198-202.

[6]. L. R. Rabiner, B. H. Juang, "An Introduction to Hidden Markov Models". IEEE Acoust. Speech Signal Proc. Magazine, Vol. 3, pp. 4-16, Jan., 25, 1986.

[7]. L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected applications in Speech Recognition" Proc. IEEE, vol. 77 (2), pp. 257-286, Feb. 1989.

[8]. J. Olsson, "Text Dependent Speaker Verification with a Hybrid HMM/ANN System". Thesis Project in Speech Technology, 2002. http://www.speech.kth.se/ctt/publications/exjobb /exjobb_jolsson.pdf.

[9]. F.Botti, A.Alexander, and A.Drygajlo. An interpretation framework for the evaluation of evidence in forensic automatic speaker recognition with limited suspect data. In *Proceedings of 2004: A Speaker Odyssey*, pages 63-68, Toledo, Spain, 2004.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 335-338

Null Values Estimation Method Based on Predictions in Incomplete Information Systems

Yanji Jiang, Ze Jiang, and Fenggang Huang
Computer Science and Technology College of Harbin Engineering University, Harbin, China
Email:790417@sina.com

***Abstract*—Collaborative filtering rating mechanism to valuation method is introduced into this paper for the shortcoming of method in improved SIM-EM (Original-EM), it proposes an improved estimation method of a null value from the perspective of the prediction of similar objects. By calculating the sparse degree of rough sets can choose suitable estimation methods. According to the degree of similarity between objects, we can predict and fill the null value when the sparsity within a specified threshold. The improved algorithm is proposed in this paper to deal with sparse data on the extreme situation can get an excellent result, and the accuracy is better than the original method.**

***Index Terms*—rough set theory; incomplete information system; null value; prediction**

I. INTRODUCTION

Rough set theory was first put forward by Polish scientists, Paw-lak [1] in 1982, it proposed a data analysis theory, which is used to reflect the capability of dealing with incomplete information or knowledge that can not distinguish between phenomena, and data classification by people, machine learning, soft computing, decision analysis, inductive reasoning [2], expert systems [3], pattern recognition, finance has been applied in many areas now. Classical rough set theory is based on incomplete information systems theory. However, in the reality the information system is usually incomplete information system, that is, there is some uncertain information in the system which is shown as a non-value data table (it contains one or more non-value). Four uncertain factors are analyzed in literature [4]: Discrete treatment, non-precise data, missing values, multiple descriptors. Among them, missing value is null value which refers to the corresponding property value unknown or unavailable. In order to make rough set theory adapt to deal with the incomplete information system , at present there are two main ways: First, an indirect approach, which is characterized by transmitting incomplete information system into incomplete information system by certain methods (often based on probability and statistics) , namely, the filling of data; second, a direct approach, which is characterized by expanding the relational concepts of the classical rough set theory in incomplete information systems, namely, the expansion of relational model. Indirect method is required to estimate null values in data pre-processing stage in order to complete the incomplete information systems. The current null value estimation methods, including special attribute value method, method on statistical analysis, model on the Bayesian forecasting, method on rough set theory. But the special attribute value method cannot reflect the nature of the empty value data and data inter-relationships easily; the method on statistical analysis cannot guarantee results or lead to a large number of calculations; Bayesian forecasting model is more complicated. Tolerance relation, asymmetrical similarity relation, limited to tolerance relation is generally used in the direct method.

Null value estimation method based on similar relation model[4] is given by literature [5] (short for SIM-EM), which gives full consideration to the compatibility of the data and the depend relationship of attributes, transitional provisional information system is avoided to generate in ROUSTID [6] algorithm, with a combination of voting strategies can get better effect of estimate. However, there are also some shortcomings in SIM-EM: ① there is no compatible class in some objects, voting strategies and other ways have to be relied on to complete the null value estimation; ②Voting strategy will fail when objects are not only in some certain compatible object, the voting strategy is actually mode method, so voting strategies will lead to failure because the problem that mode method owns "multi-mode" (two or more occurrences of the attribute value is the maximum) and "non-plural" (occurrences of all properties value are the same); ③ Not suitable for dealing with sparse data information system. The direction-area of attributes and simple-majority-rule ratio of objects are defined by literature [7] to solve ① and ② in the SIM-EM algorithm. On this basis, the algorithm in literature [8] is further improved from the view of similarity in this paper, namely the null value of property of an unknown object are predicted by the similarity between objects with similar a relationship, which instead of voting strategy, it not only overcomes the deficiencies of SIM-EM, but also is good at dealing with problem of estimation method in the extreme sparse condition.

II. RELATED CONCEPTS

A. Basic definitions

Definition 1: Information system. Given an information table $I =< U, A, V, f >$, where, $U = \{e_1, ..., e_n\}$ is a non-empty finite set of objects; A is a non-empty finite set of attributes; $V = \bigcup_{a_i \in A} V_{a_i}$ is a value set of attributes; for every $a \in A$, there is a mapping *a*, $f(x, a) \in V$, where V_a is called the value set of *a*.

Definition 2: Incomplete information system. Given an information table $I =< U, A, V, f >$, where,

$U=\{e_1,...,e_n\}$ is a non-empty finite set of objects; A is a non-empty finite set of attributes; $V=\bigcup_{a_i\in A}V_{a_i}$ is a value s et of attributes; for every $a\in A$, there is a mapping a, $f(x,a)\in V$,where V_a is called the value set of a, and at least one attribute of object is null (shown as "*") in the s ystem, denotes $\exists a(x)=*$.

Definition 3: Sparsity. It means the ratio of number of the known attributes and number of all attributes in a rough set of data tables.

Definition 4: Similarity relation. Given a subset of non-empty attributes $B\subseteq A$, S_B is a binary relation on U, for any $x, y\in U$ and $a\in B(x)\cap B(y)$, there is $x\ S_B\ y$, If and only if $a(x)=a(y)$, says S_B is a binary relation on U.

Definition 5: Compatible class. For information system $I=<U,A,V,f>$, there are any objects $x, y\in U$, if $x\ S_B\ y$, then says x and y is compatible. All the sets of objects compatible with x are called compatible class, denotes $S_B(x)$

Definition 6: Direction-area. For $\forall a\in A$ on incomplete information system $I=<U,A,V,f>$, the evaluation mid-value (denotes a_{med}) as an boundary, for $x\in U$ the area where the biggest non-empty cardinal number locates in are called direction-area, denotes $b(a)$.

Definition 7: Direction-area value. For $\forall a\in A$ on incomplete information system $I=<U,A,V,f>$, the sum of estimation value that locates in $b(a)$ is called direction-area value of a, denotes $v(b(a))$, have that:

$$v(b(a))=\sum_{x\in b(a)}\{a(x)\mid a\in A\wedge a(x)\neq\varnothing\} \quad (1)$$

Definition 8: Simple-majority-rule ratio. For $x\in U$ on $\forall a\in A\wedge a(x)\neq\varnothing$, the ratio that satisfies simple-majority-rule is called simple-majority-rule ratio of x on A:

$$\gamma(x)=\frac{card(x\mid x\in U,a\in A\wedge a(x)\neq\varnothing,meet(x,a))}{card(x\mid x\in U,a\in A,a(x)\neq\varnothing)} \quad (2)$$

Definition 9: Estimation value. The calculation formula proposed in literature [8] is as following:

$$v'(x,a)=\frac{v(b(a))}{card(x\mid x\in b(a))}\times\gamma(x) \quad (3)$$

B. Predicting value calculation

Given a known set of attribute of object A which is expressed by U_A, then U_{ij}: $U_{ij}=U_i\cup U_j\ \ (i\neq j)$.

According to the obtained set of attribute U_{ij}, three kinds of methods of similarity measure are introduced to calculate the similarity between object i and object j, in which related similarity and cosine similarity are consistent, the formulas are as following:

1) Cosine similarity formula: The property value is to be seen as a two-dimensional vector on the data sheet, if the properties of the object is unknown, then the property value is set to 0, the similarity of objects are measured by the cosine angle measure. The property value of Object i and object j on the two-dimensional vector data table respectively are expressed as $\vec{i},\vec{j}$.

$$sim(i,j)=\cos(\vec{i},\vec{j})=\frac{\vec{i},\vec{j}}{\left\|\vec{i}\right\|*\left\|\vec{j}\right\|} \quad (4)$$

The numerator is the inner product of the vector of two attribute values, the denominator is the product of the two vector module of attribute values.

2) Vector related similarity formula: The known attribute values of object i and the known attribute values of object j are presented by U_{ij}, then the $sim(i,j)$ of object i and object j are shown by Pearson related coefficient measurement formula as following:

$$sim(i,j)=\frac{\sum_{c\in Uij}(R_{i,c}-\overline{R_i})(R_{j,c}-\overline{R_j})}{\sqrt{\sum_{c\in Uij}(R_{i,c}-\overline{R_i})^2}\sqrt{\sum_{c\in Uij}(R_{j,c}-\overline{R_j})^2}} \quad (5)$$

$R_{i,c}$ is the attribute value related to object i, $\overline{R_i}$ and $\overline{R_j}$ are respectively the average attribute value of object i and object j.

3) Modified vector related similarity formula: Different objects-scale of the attribute value problems is not taken into account in the cosine similarity measurement method, the modified cosine similarity measurement method improved the above defect by subtracting the average attribute value, the $sim(i,j)$ of object i and object j is shown as formula:

$$sim(i,j)=\frac{\sum_{c\in Uij}(R_{i,c}-\overline{R_i})(R_{j,c}-\overline{R_j})}{\sqrt{\sum_{c\in Ui}(R_{i,c}-\overline{R_i})^2}\sqrt{\sum_{c\in Uj}(R_{j,c}-\overline{R_j})^2}} \quad (6)$$

This approach not only can effectively solve the problem of shortcoming of the common known attribute data, and can effectively solve the problem that all unknown attribute values are the same in cosine similarity measurement method and modified cosine similarity measurement method, it makes the calculated attribute values more accurate, thus effectively improves the quality of null value estimation. The key is how to estimate the unknown attribute values of object i in the attribute set U_{ij}. The object i in attribute set U_{ij} which attribute value is unknown is presented by N_i , namely: $Ni=Uij-Ui$

(1) The compatible class of object i is obtained by similarity relation;

(2) The neighbor object set of object i is consisted by some of objects with the highest similarity in compatible class, that is, search out a set of object $M_p=\{I_1,I_2,...,I_v\}$ in the compatible class of object i, which makes $i\in M_p$, and $sim(i,I_1)$ is the highest similarity between object I_1 and object i, $sim(i,I_2)$ is the second highest one and so on.

(3) After M_p is obtained, attribute value $\Pr e_{x,p}$ is estimated , shown as formula :

$$\Pr e_{x,p} = \overline{R_p} + \frac{\sum_{j \in M_p} sim(i,j) * (R_{j,p} - \overline{R_p})}{\sum_{j \in M_p} (|sim(i,j)|)} \quad (7)$$

$\overline{R_p}$ is the average value of attribute p in the compatible class, $R_{j,p}$ is the attribute value of p on object j.

III. IMPROVED RATING PREDICTION ALGORITHM

A. Sparsity τ and similarity weight λ

The null estimation on similarity prediction which is proposed in this paper is suitable for estimating a sparse data table, and the method on direction-area is suitable for estimating non-sparse data table, thus a new weight τ about the size of sparsity is defined to extinguish the sparse data table and non-sparse data table:

$$\tau = \frac{card(the\ known\ attribute\ values)}{card(all\ the\ known\ attribute\ values)} \quad (0 < \tau \le 1) \quad (8)$$

A suitable estimation method of data table is chosen by setting the threshold of τ ,that is, the method based on predicting value is adopted when the data table is sparse, the τ increase as the null values decrease, when the τ is big enough, the method based on direction-area keep filling up the null value. On the one hand the sparse problem of data table is solved by introducing sparsity τ , on the other hand "Non-compatible class", "multi-mode" and "non-plural" are avoided with the method based on direction-area.

According to the definition of similarity, two problems of similarity prediction are to be seen. On the one hand, the similarity of the unknown attribute is too much emphasized and the possible differences of them are ignored, for example, given an object $x = (1, *, *, *)$, $y = (1, 2, 3, 0)$, according to the known attributes, *x* and *y* have a high similarity, and other attributes merely have the possibility of the sameness, but this possibility is not so high, if join a new weight λ after obtaining the initial similarity which reduces the weight of unknown attribute, you can guarantee that the similarity of *x* and *y* is more realistic. On the other hand, the different attribute value will play a certain disrupted role in a information system which requires a higher precision, such as $x = (1, 2, 3, 0)$ and $y = (1, 2, 3, 1)$, *x* and *y* have a high similarity, but the results and the actual requirements are led to a great diversity because of the high accuracy of the system is required, so in considering the degree of similarity between objects ,it should be considered that the proportion of the unknown attributes and the different attributes, namely, the weight λ is introduced:

$$\lambda = \frac{card(attribute\ of\ x\ and\ y\ are\ the\ same)}{card(all\ attribute\ of\ x)} \quad (9)$$

The weight λ is used when there are superior attributes, which reduces the risk of estimation.

D. Algorithm description

In the Algorithm blow, step 2-12 complete the calculation of the prediction of null value, the sparsity is big enough to use the method based on direction-area in step 4 which the null value is predicted by formula (3). Step 5-12 predicts the null value according to the similarity and similarity weight when the data table is sparse. Step 13-14 fill the correspondingly null value by predicting value getting from step 4 or step 5-12, and then loop to the next prediction. The next incomplete object is going to be filled when there are not null values in the current object.

Algorithm: NVP (Null Value Prediction)
Input: incomplete information system I
Output: complete information system I'
(1) for exit incomplete object *i* in I
(2) for exit null attribute value in object *i*
(3) τ = card (non-null attribute value)/ card (all attribute value);
(4) if $0.5 < \tau \le 1$ then $\Pr e(i,p) = v(i,p)$;
(5) if $0 < \tau \le 0.5$ then {
(6) calculate $S_B(i)$;
(7) for $j \in S_B(i)$ and $j \ne i$
(8) calculate $sim(i,j)$;
(9) $\lambda_j = card(a(i) = a(j)) / card(a(i))$;
(10) next *j*
(11) order $I_j = sim(i,j) \times \lambda_j$ from big to small as $M_p = \{I_1, I_2, ..., I_v\}$;
(12) calculate $\Pr e(i,p)$; }
(13) if $\Pr e(i,p) \ne \varnothing$ then $p(i) = \Pr e(i,p)$; else $p(i) = *$;
(14) next *p*
(15) next *i*

IV. IMPROVED RATING PREDICTION ALGORITHM

The incomplete information table of literature [10] is introduced in this paper, which contains the range of $(1, 2, 3, 4)$ the attribute of (a_1, a_2, a_3, a_4) , $0.5 < \tau \le 1$ in the data table is shown in table 1. A non-null value is chosen from the data table to replace a null value, Original-EM and Proposed-EM are used to estimate the chosen element respectively. In order to verify the validity of the method proposed in this paper, table 1 is converted to table 2 which $0 < \tau \le 0.5$ is satisfied, also Original-EM and Proposed-EM are used to estimate the chosen element respectively.

TABLE 1. INCOMPLETE DATA TABLE

A	a1	a2	a3	a4
x1	3	2	1	0
x2	2	3	2	0
x3	2	3	2	0
x4	*	2	*	1
x5	*	2	*	1
x6	2	3	2	1
x7	3	*	*	3
x8	*	0	0	*
x9	3	2	1	3
x10	1	*	*	*
x11	*	2	*	*
x12	3	2	1	*

TABLE 2. INCOMPLETE SRASITY DATA TABLE

A	a1	a2	a3	a4
x1	3	*	*	0
x2	*	3	*	0
x3	2	*	*	0
x4	*	2	*	1
x5	*	2	*	1
x6	2	3	2	*
x7	3	*	*	3
x8	*	0	0	*
x9	3	2	*	*
x10	1	*	*	*
x11	*	2	*	*
x12	*	2	1	*

TABLE 3. ESTIMATION OF OTIGINAL-EM AND PROPOSED-EM ON SPARSE TABLE

	Actual Value	Original-EM	Proposed-EM
v(x1,a1)	1	0	1
v(x1,a4)	0	1	0
v(x2,a2)	3	3	3
v(x2,a4)	0	0	1
v(x3,a1)	2	2	2
v(x3,a4)	0	0	0
v(x4,a2)	2	1	1
v(x5,a2)	2	3	2
v(x5,a4)	1	1	1
v(x6,a1)	2	1	2
v(x6,a2)	3	3	3
v(x6,a3)	2	1	1
v(x7,a1)	3	2	2
v(x7,a4)	3	2	2
v(x8,a2)	0	1	0
v(x8,a3)	0	0	0
v(x9,a1)	3	3	3
v(x9,a2)	2	2	2
v(x10,a1)	1	0	0
v(x11,a1)	3	1	3
v(x12,a2)	2	2	2
v(x12,a3)	1	0	0

The evaluation criteria of accuracy rating is used to compare Original-EM algorithm and improved algorithm Proposed-EM).

Accuracy rating: it refers to the ratio of the total number of correct estimation attribute values and the total number of non-null attribute values, denotes *C*:

$$C = \frac{card(x \mid x \in U, a \in A, a(x) \neq \varnothing \wedge \Pr e(x,a) = a(x))}{card(x \mid x \in U, a \in A, a(x) \neq \varnothing)} \quad (10)$$

As shown in table 1, sparsity $\tau = 0.68$, method based on direction-area is used by Original-EM and Proposed-EM, because that the data table is a non-sparse data table, the accuracy rating are the same now, $C_O = C_P = 66.7\%$. As shown in table 2, sparsity $\tau = 0.45$, the data table is sparse, as shown in table 3, the result is estimated by Original-EM and Proposed-EM respectively. Accuracy rating C_O=45.5% , $C_P = 63.6\%$ now. The accuracy rating of Proposed-EM is as 18.1% higher than Original-EM. Thus, in the estimation of sparse data table, the algorithm is presented in this paper more effective.

IV. CONCLUSION

Applying rough set theory into incomplete information system is one of the key to propel it into practical, because the data shows that the data which needs to be addressed is a certain degree of incomplete. For the problem of null value estimation in incomplete information system, the literature [8] gives a method based on direction-area to the estimate null values, through a combination of simple-majority-rule, it can get better effect, but there is also a lack of diversity of data processing problems. In this paper, an improved method is proposed from view of the predictive value on collaborative filtering; sparsity and similarity weight are defined to combine with attribute predicting value, in this way, the quality of estimation method is ensured in both cases of dense data and sparse data. The accuracy rating is superior than the original method.

REFERENCES

[1] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982,11(5):341-356.

[2] SHENG Bu-yun, LIN Zhi-jun, DING Yu-feng, Rough set-based conflict resolution case reasoning in collaborative design [J].Computer Integrated Manufacturing Systems, 2006,12(12):1952-1956.

[3] XUE Jun-fang, XIANG Dong, QIU Chang-hua. Knowledge acquirement method for component integration based on rough set[J]. Computer Integrated Manufacturing Systems ,2007,13(8):1658-1664.

[4] SLOWINSKI R, STEFANOWSKI J. Handing various types of uncertainty in the rough set approach[C]//Proceedings International Workshop on Rough Sets and Knowledge Discovery: Rough Sets, Fuzzy Sets and Knowledge Discovery. London, UK: Springer-Verlag, 1993:366-376.

[5] YANG Shan-lin. Intelligent decision-making methods and Intelligent Decision Support System [M].Beijing: Science Press,2005.

[6] SLOWINSKI R, VANDERPOOTEN D. A generalized definition of rough approximations based on similarity [J]. IEEE Transactions on Knowledge and Data Engineering, 2000, 12(2):331-336.

[7] MENG Jun, LIU Yong-chao, Mo Hao-bo. New method of packing missing data based on rough set theory. Computer Engineering and Applications.2008,44(6).

[8] LI Cong, LIANG Cang-yong, YANG Shan-lin. Null values estimation method based on rough set for incomplete information systems. Computer Integrated Manufacturing Systems. 2009.3:Vol.15,No 3.

[9] GUO Yan-hong. On Clooaborative Filtering Algorithm and Applications of Recommender Systems. Dalian University of Technology , PHD thesis.2008.4.

[10] WANG Guo-yin. Extension of rough set under incomplete information systems, Journal of Computer Research and Development.2002.10,Vol.39,No.10.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 339-342

An Extension of WSDL for Flexible Web Service Invocation with Large Data

Aihua Wu
Dept. of Computer Science
Shanghai Maritime University, Shanghai, China
ahwu@cie.shmtu.edu.cn

***Abstract*--Traditional web services are limited in the request-reply framework, where service invocation can only be bundled together with all relevant data in a single message. However, more and more services require large datasets (e.g., multimedia and scientific data). Under the WSDL standards, these datasets must be ported to an appropriate message format and transferred in their entirety upon each service invocation or response. A strategy to solve this problem which is called WSDL-D [1] is to separate invocation messages from their datasets. Such a separation ultimately grants service consumers the ability to pass parameter datasets from third party hosts, to maintain dataset parameters on the service provider host for use with future service invocations, and to provide datasets in a variety of different formats. In this paper, we extend the language of WSDL so that service invocation mechanism of WSDL-D can be possible.**

Index term---**WSDL, web service, service invocation, large dataset**

1. INTRODUCTION

As web technologies advance and the range of applications utilizing web services expands, a need for web service invocation management that can efficiently handle large datasets arises. For example, in scientific workflow applications, both algorithms and datasets (usually large in size) are frequently published and shared [4, 5, 6, 7]. Also, third party data management is becoming a new and interesting model for data management and IT operations for businesses. On the other hand, the design of WSDL was primarily based on the motivation to loosen coupling of interoperating software components in a simple framework. While this is achieved in WSDL and useful in many non-data intensive applications, the appearance of large datasets is causing many efficiency problems in the WSDL framework.

Current WSDL web services are built around the request-reply framework, requiring service invocation messages to include all relevant data. For lightweight web services, it is appropriate. However, when massive datasets (e.g., multimedia data, scientific datasets) need to be transferred between user and the server, it becomes inefficiency. Under the current WSDL and related standards, these large datasets must be ported to an appropriate message format and transferred in their entirety upon each service invocation or response.

Approach of WSDL-D [1] makes significant improvements in service flexibility and performance stand simply by separating invocation and response messages from their respective datasets. Such a separation not only grants service consumers the ability to pass parameter datasets from third party hosts, but also to maintain dataset parameters on the service provider host for use with future service invocations, and to provide datasets in a variety of different formats.

In this paper, we extend the standard of Web Service Description Language so that it can support service invocation mechanism of WSDL-D which separates service invocation and dataset transferring between service caller and service provider. In WSDL-D, input parameters (datasets) of a service invocation is not required to be sent with the invocation message, instead, an input dataset can be fetched by a service provider, or sent later by the requester, or simply the dataset used in previous invocation requests. Similarly, an output dataset can be pushed to the requester asynchronously, or fetched by the requester. However, there is no mechanism to support complex input and output parameters required in WSDL-D.

This paper makes the following technical contributions: (1) explore of issues in traditional web service invocation when large dataset need to be transferred, and (2) syntax augmentation for WSDL which is compatible with WSDL.

This paper is organized as follows. Section 2 illustrates application scenarios to motivate the dataset problems in WSDL. Section 3 presents details of the design and extension of WSDL. Section 4 concludes the paper.

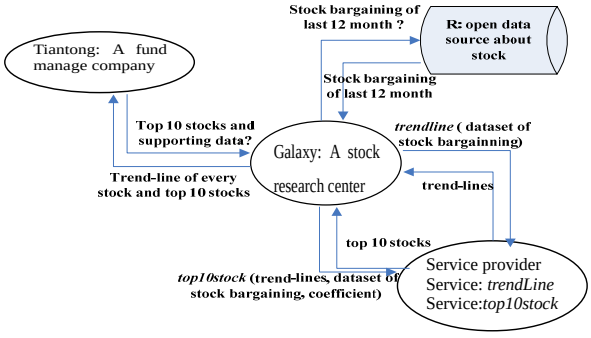
2. PROBLEM AND MOTIVATION

Traditional web service call is a request and response communication between service user and service provider. Communication patterns here are limited to one-way, request-response, solicit-response and notification. Typically, data related to the invocation is small. It works well in some applications, but there are still many applications require more than that.

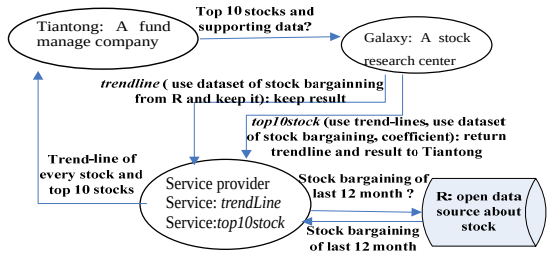
Scenario Just like (a) of figure 1 shows , as a fund management company, *Tiantong* wants to know *Galaxy*'s (a stock research center) opinion about what is the top 10 stocks deserved to invest in the next month and also supporting data for this conclusion. To give answer to this question, *Galaxy* need each stock's price trend line

and stock bargaining information of the last 12 month, and a coefficient. The coefficient is stored in *Galaxy*, stock bargaining information of the last 12 month, with which web service *trendLine* can calculate trend line for each stock, can be obtained from an open data source published by the stock exchanger. And given the three information, web service *top10stock* can tell Galaxy which is the top 10 stocks deserved to invest in the next month.

In traditional way, as (a) shows, to response to *Tiantong*'s request, *Galaxy* will: 1) query and download data about last 12 month stock bargaining (denoted as *SBargaining*) from R; 2) invocate service trendline with *SBargaining*; 3) after receiving trend lines of all stock, invocate *top10stock* with *SBargaining*, set of trend lines and coefficient; 4) after receive the result, send it and the set of trend lines back to *Tiantong* as the answer.



(a) Traditional web service invocation



(b) One new invocation

Figure1: an example where more than two parts are involved in web service

First, all dataset is wrapped in XML file in traditional way. Service *trendLine* can not start until all *SBargaining* arrived, and if the XML file failed in transferring, the invocation failed. In fact, when the first stock bargaining data arrive, *trendLine* can start calculating its trend line while waiting for the others at the same time. And even part of *SBargaining* are failed in transferring, there is no need to retransfer the whole *SBargaining*.

Second, notice that during the whole process, both *SBargaining* and trend lines are transferred three times.

There are thousands upon thousands stocks, both *SBargaining* and trend lines are big datasets. Obviously, frequent transfer of large datasets will cause the inefficiency of the whole process.

Suppose R, the data center, and *Tiantong*, the actual data consumer, can be a part of the web service framework, as (b) shows, duplicate data transfer will be reduced. In (b), *Galaxy* doesn't get data from R, instead, he tells service provider to get *SBargaining* from R and keep it for future use when he invocate *trendLine*, he also requires keeping the result for future use. Similarly, Galaxy tells service provider to reuse SBargaining and trend lines when he invocate *top10stock*, and return the result and trend lines directly to *Tiantong*. Obviously, (b) is more efficient.

From the above scenario, traditional web service framework is not excellent in or misses something in some complicated application:

- When input parameters and output result of web service is large dataset, how to deal with the invocation efficiently? How to deal with other format input data, for example binary file?
- When input data isn't stored in service user but in a data center, or when output data is not wanted by service user but by the third part, how to deal with the service invocation efficiently?
- Can service provider get data from the service user instead of waiting for service user sending to him?
- When data can be reused, data transfer executor and data consumer can be a third part, how to change the communication pattern between them?

In framework of WSDL-D, third or more part can be involved, back and forth communication can happen between service user and service, large dataset can be dealt efficiently and data can be reused. To support all these flexible invocation, we extend the standard language of WSDL in this paper.

3. Syntax Extension of WSDL

In WSDL-D applications, when service requestor invocated a service, what he needs to do is not only telling service parameter content but also features of that parameter. Such feature can be specified by some attributes of the parameter. In general, such attribute include:

- Data transfer method from service requestor to service provider, which we named *transferMethodIn*. Possible *transferMethodIn* can be *clientPush* which means that dataset corresponding to this message will be pushed to the service provider by service requestor, *serverPull* which means that dataset will be pulled from the service requestor by service provider, and *serverUse* which means that no data will be transferred but service should reuse an existing data.
- Data transfer method from service provider to service requestor, which we named *transferMethodOut*. Possible *transferMethodOut* can be *clientPull* which means that dataset corresponding to this message will be pulled from the service provider by service requestor, and *serverPush* which means that dataset will be pushed to service requestor by service provider.
- Address of dataset related to the message type, which we named *address*. So that data transfer executor know where to get the data.
- Exiting data ID, which we named *dataID*, used in

input message. When *transferMethodIn* is *serverUse*, this attribute can telling service provider which existing data should be used.

- Attribute *persistent.* By this attribute, service requestor can ask service provider to return id of dataset related to this message so that this dataset can be reused in the following invocation.

Besides these attribute,

- To be compatible to WSDL file, WSDL-D file must allow sub element in the message type can be a traditional web service parameter with value of basic XML data type and also can be an element with some of the above attribute but with no value. Thus, the attribute which is nillable of this sub element should be true.
- To allow service requestor define some features of the service output, WSDL-D files define an optional element for each output parameter in input message, accordingly, service provider can know that in returning result, which data transfer method should take, where result should be send to, and whether persistent store the result.

All input or output XML files based on WSDL-D must conform to the following extended constraint:

a) When any sub element of input or output message is not empty, all extended attribute must be ignored.

b) When any sub element of input or output message is empty, its *transferMethodin* or *transferMethodOut* can not be ignored. And: 1) when *transferMethod* is *serverPull* or *clientPull*, its address can not be ignored; 2) When *transferMethodIn* is *serverUse*, its *dataID* can't be ignored.

c) When element resultFeature exist: 1) if its subelement persistent is true, attribute *dataID* of the related output part can not be ignored. 2) Subelement *transferMethodOut* of *resultFeature* and attribute *transferMethodOut* of the output part *resultFeature* should share the same value.

d) There can not exist a *resultFeature* refer to a pure WSDL output part whose type definition doesn't include attribute *transferMethodOut*.

Example The following is a WSDL-D file and input/output XML file based on it.

```
……
<wsdl:types>
  <s:schema elementFormDefault="qualified"
      targetNamespace="http://tempuri.org/">
    <s:simpleType name = "transferIn">
      <s:restriction base = "xsd:string">
      <s:enumeration value = "clientPush"/>
      <s:enumeration value = "serverPull"/>
      <s:enumeration value = "serverUse"/>
    </s:simpleType>
    <s:simpleType name = "transferOut">
      <s:restriction base = "xsd:string">
      <s:enumeration value = "clientPuLL"/>
      <s:enumeration value = "serverPush"/>
    </s:simpleType>
    <s:element name="HelloWorld">
      <s:complexType>
        <s:sequence>
          <s:element minOccurs="1" maxOccurs="1" name="x"
              nillable="true" type="s:int">
          <s:attribute minOccurs="0" maxOccurs="1"
              name="transferMethodIn" type="tns:TransferIn"/>
          <s:attribute minOccurs="0" maxOccurs="1"
              name="address" type="s:URI"/>
            <s:attribute minOccurs="0" maxOccurs="1"
                name="persistent" type="s:boolean"/>
            <s:attribute minOccurs="0" maxOccurs="1"
                name="dataID" type="s:string"/>
      </s:element>
      <s:element minOccurs="1" maxOccurs="1" name="y" type="s:int"/>
      <s:element minOccurs="0" name="resultFeature" ref="">
       <s:element minOccurs="0" maxOccurs="1"
          name="transferMethodOut" type="tns:TransferOut"/>
       <s:element minOccurs="0" maxOccurs="1" name="address"
          type="s:string"/>
      <s:element minOccurs="0" maxOccurs="1"
          name="persistent" type="s:boolean"/>
      </s:element>
     </s:sequence>
    </s:complexType>
   </s:element>
   <s:element name="HelloWorldResponse">
    <s:complexType>
     <s:sequence>
      <s:element minOccurs="1" maxOccurs="1" name="z"
nillable="true"
        type="s:int">
        <s:attribute minOccurs="0" maxOccurs="1"
            name="transferMethodOut" type="tns:TransferOut"/>
        <s:attribute minOccurs="0" maxOccurs="1" name="address"
            type="s:string"/>
        <s:attribute minOccurs="0" maxOccurs="1" name="dataID"
            type="s:string"/>
       </s:element>
      </s:sequence>
     </s:complexType>
    </s:element>
   </s:schema>
  </wsdl:types>
<wsdl:message name="HelloWorldSoapIn">
      <wsdl:part name="parameters" element="tns:HelloWorld" />
 </wsdl:message>
<wsdl:message name="HelloWorldSoapOut">
     <wsdl:part name="parameters" element="tns:HelloWorldResponse" />
 </wsdl:message>
<wsdl:portType name="Service1Soap">
  <wsdl:operation name="HelloWorld">
              <wsdl:input message="tns:HelloWorldSoapIn" />
              <wsdl:output message="tns:HelloWorldSoapOut" />
  </wsdl:operation>
</wsdl:portType>
<wsdl:binding name="Service1Soap" type="tns:Service1Soap">
 <soap:binding transport="http://schemas.xmlsoap.org/soap/http" />
  <wsdl:operation name="HelloWorld">
    <soap:operation soapAction= "http://tempuri.org/HelloWorld"
    style="document" />
              <wsdl:input>  <soap:body use="literal" />  </wsdl:input>
               <wsdl:output> <soap:body use="literal" />  </wsdl:output>
  </wsdl:operation>
```

```
 </wsdl:binding>
<wsdl:service name="Service1">
  <wsdl:port name="Service1Soap" binding="tns:Service1Soap">
     <soap:address location="http://localhost:2465/Service1.asmx" />
  </wsdl:port>
 </wsdl:service>
</wsdl:definitions>
```

(a)WSDL-D definition

```
……
 <soap:Body>
  <HelloWorld xmlns="http://tempuri.org/">
    <x transferMethodIn=”serverPull”
       address=”http://tempuri.org/data/hw.bin” persistent=”true” />
    <y>10</y>
    <resultFeature >
      <transferMethodOut>serverPush</transferMethodOut>
      <address> http://www.ucsb.edu/data </address>
      <persistent> true </persistent>
    </resultFeature>
  </HelloWorld>
 </soap:Body>
</soap:Envelope>
```

(b) One valid input XML file

```
……
 <soap:Body>
  <HelloWorldResponse xmlns="http://tempuri.org/">
    <z transferMethodOut=”serverPush” address=”a.bin”
       dataID=”109se9r6t57”/>
  </HelloWorldResponse>
 </soap:Body>
</soap:Envelope>
```

(c) One valid output XML file

Figure 2. an example of our extension to WSDL

In the example, we can learn from the input XML file that two parameters are required to call service: *y*, a simple integer number, and *x* is a large dataset who inhabits in http://tempuri.org/data/ (may be the third party) with name of *hw.bin*. The server is responsible to pull it out and store it for future reuse. As for the result, according to the input file, it will be pushed by the server to and persistently stored on http://www.ucsb.edu/data (may be the third party).

From the output XML file, we know that the result who is also a large dataset with name of a.bin will be pushed by the server to host required by the input file. Here a.bin is assigned with an id “109se9r6t57” so that it can be reused without being transferred when it is the parameter of another call of the service in the future (data reuse).

From the example, it is obvious that our extension can support large dataset, data reuse and third party.

4. Conclusions

WSDL provides a simple interface for web services. However, it lacks support for handling large datasets. This paper presents an extension of WSDL to allow decoupling the invocation request and invocation parameters (datasets). The extension allows, e.g., obtaining datasets from a third party, reusing a prior dataset, etc.

5. References

[1]Mark Wiley, Aihua Wu, Jianwen Su: WSDL-D: A Flexible Web Service Invocation Mechanism for Large Datasets. CEC/EEE 2008: 157-164

[2] R. Akkiraju, J. Farrell, et al. Web Service Semantics - WSDL-S, W3C Member Submission, November 7, 2005 (http://www.w3.org/Submission/WSDL-S/)

[3] D. Box, L.F. Cabrerea, et al. “Web Services Eventing (WS-Eventing)”, W3C, March 15, 2006 (http://www.w3.org/Submission/WS-Eventing/)

[4] Earth System Grid, http://www.earthsystemgrid.org/

[5] Federation of Earth Science Information Partners, http://www.esipfed.org

[6] B. Ludäscher and C.A. Goble. Guest editors’ introduction to the special section on scientific workflows. SIGMOD Record, 34(3):3-4, 2005

[7] C. Reed, “Integrating Geospatial Standards and Standards Strategies into Business Processes”, 2004 (available from http://www.opengeospatial.org/pressroom/papers)

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 343-348

The Orthogonal Decomposition Algorithm for Speech Signals in Reproducing Kernel Space

Sen Zhang, Lei Liu, and Luhong Diao
College of Applied Sciences, Beijing University of Technology, Beijing, China
Email: zhangsen@yahoo.com

***Abstract*—An orthogonal decomposition method and implementation algorithm for speech signal processing are proposed in this paper. In the reproducing kernel function of Hilbert space $W_2^1[a,b]$, a set of normalized orthogonal function system $\{\varphi_j^*(x)\}_1^n$ is generated, and speech signals can be orthogonally decomposed in $W_2^1[a,b]$ according to the basis $\{\varphi_j^*(x)\}_1^n$, the orthogonal decomposition coefficients can be computed by a fast algorithm based on the properties of reproducing kernel function. This approach mapped the speech signals represented by discrete samples to continuous functions which is different from the canonical form represented by series of triangle functions, and the inner product computation in Hilbert space was transformed into function evaluation problem only at some discrete points.**

***Index Terms*—Speech Signal, Reproducing Kernel Space, Orthogonal Decomposition, Signal Analysis**

I. INTRODUCTION

The orthogonal decomposition methods (ODM) have been widely applied in speech signal analysis and processing fields. By orthogonal decomposition, speech signals can be projected into a special subspace and the projections or coefficients can be used to represent the speech signals' features which could be used in speech coding to compress the redundancy, etc. As we know, Fourier transformation and wavelet transformation are typical orthogonal decomposition methods which have been successfully utilized in speech signal processing fields and other engineering fields. However, the disadvantages of these orthogonal decomposition methods have been thoroughly investigated and were proposed in many research reports [16].

In recent years, Reproducing Kernel (RPK) space theory has been used in some engineering fields, such as classification methods based on Support Vector Machine (SVM), pattern recognition, machine learning, image compression and reconstruction[1-6], and these successful applications in other fields encouraged researchers to study whether and how RPKS could be used in speech signal analysis and processing fields. As a result, we know the discrete speech signals can be represented in continuous form by RPK which could reduce the computation load in speech analysis process, and the RPK operator eigenvalue theory was used to investigate the nonlinear features and the relations between the nonlinear features and the high order harmonies of speech signals[7-15]. Besides, the RPK was also used in speech signal compression and reconstruction based on the property of RPK which can produce uniform approximation for speech signals by using almost optimal interpolation operator of RPK [2,3,6]. The speech features could be also extracted in RPK space and applied in speech recognition. Some research [7-9] reported that the speech features extracted in RPK space were more robust compared to MFCC which was widely used in speech recognition today.

This paper includes six sections which are arranged as follows: firstly, the up-to-date applications of RPK theory in speech signal processing fields are introduced; next, a special RPK space $W_2^1[a,b]$ is discussed; thirdly, some problems related to orthogonal decomposition of speech signals in $W_2^1[a,b]$ are investigated; fourthly, the orthogonal decomposition approaches and algorithms of speech signals in $W_2^1[a,b]$ are proposed in details; the fifth section presents some experimental results and the evaluation, and some conclusions are given in the last section.

II. RPKS $W_2^1[a,b]$

To begin with, some fundamental results about space $W_2^1[a,b]$ will be discussed in this section[16]. Suppose a special functional space $W_2^1[a,b]$ contains some real or complex functions defined in interval [a, b] as its elements. By properly defining inner product and norm, $W_2^1[a,b]$ could become a Hilbert space, and it is also a RPK space. The RPK of $W_2^1[a,b]$ could be represented in some simple function form.

Definition 2.1 $W_2^1 \equiv W_2^1[a,b] = \{f(x) \mid f(x)$ is an absolutely continuous function defined in $[a, b]$, and its derivative function $f'(x) \in L^2[a, b]\}$, where $L^2[a, b]$ is a function set which contains all the functions which are integrable in squared form.

Definition 2.2 The inner product and norm in space $W_2^1[a,b]$ were defined respectively as follows:

$$(f,g)=\int_a^b (f(x)g(x)+f'(x)g'(x))dx,$$
$$\|f\|=(f,f)^{1/2} \qquad (2.1)$$

Theorem 2.1[16] The function space $W_2^1[a,b]$ is a complete inner product space with the inner product defined by formula (2.1).

Theorem 2.2[16] The inner product space $W_2^1[a,b]$ has a unique RPK function $K(x, y)$ which can be represented as follows:

$$K(x,y)\equiv K_x(y)=\frac{1}{2sh(b-a)}[ch(x+y-b-a)+ch(|x-y|-b+a)] \qquad (2.2)$$

Thus, $W_2^1[a,b]$ is a RPK space with RPK function $K(x, y)$.

The RPK function $K_x(y)$ of $W_2^1[a,b]$ has the reproducing property, that is

$\forall f(x)\in W_2^1[a,b]$ and $\forall y\in[a,b]$, we have

$$(f(x),K_y(x))=f(y) \qquad (2.3)$$

Next, a complete function system in space $W_2^1[a,b]$ will be constructed. Suppose the set $T=\{t_1,t_2\cdots\}$ is dense in the interval $[a,b]$, and for each $t_i\in T$, according to the property of the RPK function $K_x(y)$, we have a function $K_{t_i}(x)$, and for simpleness, we rewrite $K_{t_i}(x)$ as $\varphi_i(x)$, namely $\varphi_i(x)=K_{t_i}(x)$. Then we have the following theorem 2.3.

Theorem 2.3 The function system $\{\varphi_i(x)\}_{i=1}^{\infty}$ is a complete function system in space $W_2^1[a,b]$.

Proof: Suppose function $f(x)\in W_2^1[a,b]$.

If $(f(x),\varphi_i(x))=0, i=1,2,\cdots$, note the definition of $\varphi_i(x)$ and its reproducing property,

then $f(t_i)=(f(x),\varphi_i(x))=0, i=1,2,\cdots$.

Since the set $T=\{t_1,t_2\cdots\}$ is dense in interval $[a,b]$, and $f(x)$ is continuous in $[a,b]$, it is easy to know that $f(x)\equiv 0$, which means the function system $\{\varphi_i(x)\}_{i=1}^{\infty}$ is a complete function system in space $W_2^1[a,b]$. □

Furthermore, take $\{\varphi_i(x)\}_{i=1}^{\infty}$ to be orthogonalized through *Schmidt* approach, a standard orthogonal base $\{\varphi_i^*(x)\}_{i=1}^{\infty}$ in $W_2^1[a,b]$ is obtained, that is

$$\varphi_i^*(x)=\sum_{j=1}^{i}\beta_{ji}\varphi_j(x) \qquad (2.4)$$

where β_{ji} is the orthogonal coefficient.

Theorem 2.4 Suppose $f(x)$ is a function in $W_2^1[a,b]$, then the Fourier series of $f(x)$ regarding to $\{\varphi_i^*(x)\}_{i=1}^{\infty}$ is convergent, and we have the following formula:

$$f(x)=\sum_{i=1}^{\infty}\sum_{j=1}^{i}\beta_{ji}f(t_j)\varphi_i^*(x) \qquad (2.5)$$

Proof: Since the standard orthogonal base $\{\varphi_i^*(x)\}_{i=1}^{\infty}$ in $W_2^1[a,b]$ is complete, function $f(x)\in W_2^1[a,b]$ can be decomposed as follows:

$$f(x)=\sum_{i=1}^{\infty}(f(x),\varphi_i^*(x))\varphi_i^*(x)$$

based on formula (2.4), we obtained $f(x)$ as follows:

$$\begin{aligned} f(x)&=\sum_{i=1}^{\infty}(f(x),\sum_{j=1}^{i}\beta_{ji}\varphi_j(x))\varphi_i^*(x)\\ &=\sum_{i=1}^{\infty}\sum_{j=1}^{i}\beta_{ji}(f(x),\varphi_j(x))\varphi_i^*(x)\\ &=\sum_{i=1}^{\infty}\sum_{j=1}^{i}\beta_{ji}f(t_j)\varphi_i^*(x)\end{aligned}$$

□

We write $\alpha_i=\sum_{j=1}^{i}\beta_{ji}f(t_j)$ and call α_i generalized Fourier coefficients, where β_{ji} is the *Schmidt* orthogonal coefficient. Obviously, the computation of α_i only requires to evaluate $f(x)$ at some fixed points (e.g., t_j). In general, the computation of Fourier coefficients requires to calculate integral which usually is a heavy computation load. Comparatively, the computation of generalized Fourier coefficients in $W_2^1[a,b]$ requires much less computation cost.

Theorem 2. 5 The projection of function $f(x)\in W_2^1[a,b]$ in subspace $S=span\{\varphi_1,\varphi_2\cdots\varphi_n\}$ is represented as follows:

$$\overline{f}_n(x)=(P_nf)(x)=\sum_{i=1}^{n}\sum_{j=1}^{i}\beta_{ji}f(t_j)\varphi_i^*(x) \qquad (2.6)$$

Moreover, we have $f(t_i)=\overline{f}_n(t_i)$, and $\overline{f}_n(x)$ is uniformly convergent to $f(x)$ while $n\to\infty$, where $P_n: W_2^1[a,b]\to S$ is projection operator.

Proof: Since P_n is projection operator, it is easy to obtain the following formula:

$$(P_n f)(x)=\sum_{i=1}^{n}(f(x),\varphi_i^*(x))\varphi_i^*(x)$$

Note that $\varphi_i^*(x)=\sum_{j=1}^{i}\beta_{ji}\varphi_j(x)$ and the property of RPK, then we can obtain:

$$(P_n f)(x)=\sum_{i=1}^{n}\sum_{j=1}^{i}\beta_{ji}f(t_j)\varphi_i^*(x)$$

On the other hand, since P_n is a conjugated operator, it satisfies the following conditions:

$$\begin{aligned}\overline{f}_n(t_i)&=(P_n f)(t_i)=((P_n f)(x),\varphi_i(x))\\&=(f(x),(P_n\varphi_i)(x))=(f(x),\varphi_i(x))=f(t_i)\end{aligned}$$

Next, we will prove that $\overline{f}_n(x)$ is uniformly convergent to $f(x)$.

From formula (2.3) we know:

$$\begin{aligned}&\left|\overline{f}_n(x)-f(x)\right|=\left|(\overline{f}_n(.)-f(.),K_x(.)\right|\\&\le\left\|\overline{f}_n(.)-f(.)\right\|\left\|K_x(.)\right\|=K_x(x)\left\|\overline{f}_n(.)-f(.)\right\|\\&=\frac{ch(2x-b-a)+ch(b-a)}{2sh(b-a)}\left\|\overline{f}_n()-f(.)\right\|\\&\le M\left\|\overline{f}_n(.)-f(.)\right\|\end{aligned}$$

and based on Th.2.4, we can obtain that $\left\|\overline{f}_n(\cdot)-f(\cdot)\right\|\to 0$

thus, $\overline{f}_n(x)$ is uniformly convergent to $f(x)$.

□

Theorem 2.5 indicates that the projection of function $f(x)$ in space $W_2^1[a,b]$ is one kind of interpolation operation of $f(x)$. Furthermore, we know that projection approximation is optimal in some sense, and the interpolation operation of $f(x)$ ensures that interpolating function (the projection of function $f(x)$) equals to the original function $f(x)$ at all interpolation points. So, the computation of the projection of function $f(x)$ can be realized in space $W_2^1[a,b]$ with much less calculation load.

Based on theorem 2.4 and 2.5, we can obtain that the orthogonal decomposition of $f(x)$ in space $W_2^1[a,b]$ regarding to finite orthogonal function set $\{\varphi_i(x)\}_i^n$ is as follows:

$$\begin{aligned}f(x)&=\sum_{i=1}^{n}(f(x),\varphi_i^*(x))\varphi_i^*(x)=\sum_{i=1}^{n}\sum_{j=1}^{i}\alpha_i\beta_{ji}\varphi_j(x)\\&=\sum_{i=1}^{n}\sum_{j=1}^{i}\sum_{k=1}^{i}\beta_{ji}\beta_{ki}f(x_j)\varphi_k(x)\qquad(2.7)\end{aligned}$$

The above formula (2.7) indicates that function $f(x)$ can be denoted by finite samples $\{f(x_j)\}_1^n$ in interpolation form. In practical application, we do not require to compute all the coefficients β_{ji}. In fact, only about 3*n β_{ji} is enough.

The 3-dimensional and 2-dimensional images of the RPK function $K(x, y)$ of $W_2^1[a,b]$ were shown in the following Fig. 2.1 and Fig. 2.2, where $x, y\in[0,1]$, and y was sampled at 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.9, etc.

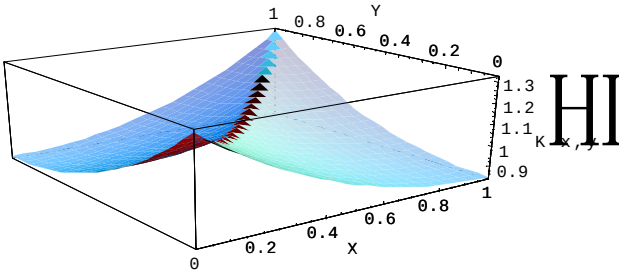


Fig. 2.1 The 3-dimensional image of $K(x, y)$

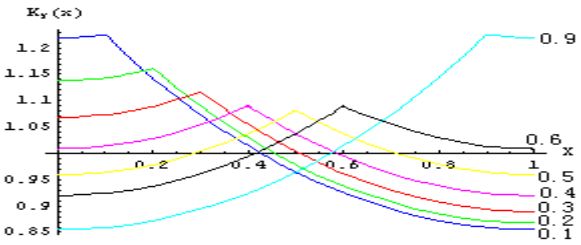


Fig. 2.2 The 2-dimensional image of $K(x, y)$

III. Decomposition in $W_2^1[a,b]$

Suppose speech signal samples were represented as f_1, f_2, …, f_n. We know from above formula (3.1), the speech signal can be decomposed in space $W_2^1[a,b]$, the orthogonal decomposition coefficients $\{\alpha_j\}_1^n$ can be computed by following formula:

$$\alpha_j=\sum_{k=1}^{j}\beta_{kj}f_k$$

where β_{kj} is the orthogonal coefficient of $\{\varphi_j(x)\}_1^n$ through *Schmidt* approach. In general, the computation of β_{kj} requires inner product computation, which usually needs to calculate integral. It is obvious that the key problem of speech signal decomposition in space

$W_2^1[a,b]$ is how to calculate β_{kj} quickly by using the properties of RPK space $W_2^1[a,b]$.

From the definition of $\{\varphi_j(x)\}_1^n$, it is a linearly independent function set and has the reproducing property. The *Schmidt* orthogonalizing process of basis functions $\{\varphi_j(x)\}_1^n$ is as follows:

$$\begin{cases} \psi_1(x)=\varphi_1(x), \quad \varphi_1^*(x)=\psi_1(x)/\|\psi_1\| \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \psi_k(x)=\varphi_k(x)-\sum_{j=1}^{k-1}(\varphi_k(x),\varphi_j^*(x))\varphi_j^*(x) \\ \varphi_k^*(x)=\psi_k(x)/\|\psi_k\|, \quad k=2,...,n \end{cases} \tag{4.1}$$

From above process we know $\varphi_j^*(x)$ is linear combination of $\{\varphi_i(x)\}_1^j$, and the computation of the inner product $(\varphi_k(x),\varphi_j^*(x))$ can be transformed into the computation of $(\varphi_k(x),\varphi_j(x)), j=1,...,k-1$. Besides, to compute $\|\psi_k\|$ requires to compute $(\varphi_k(x),\varphi_k(x))$ first. Hence, the computation of $(\varphi_k(x),\varphi_j(x)), j=1,...,k$ is necessary.

From the properties of $W_2^1[a,b]$ and the definition of $\{\varphi_j(x)\}_1^n$, it is easy to obtain the following:

$(\varphi_k(x),\varphi_j(x))=\varphi_k(x_j), j=1,...,k$ 。

Therefore, we can compute the orthogonal coefficient β_{kj} according to the following algorithm:

(a) for k=1 and j=1, compute $\|\psi_1\|$, β_{11} as:

$$\|\psi_1\|=[\varphi_1(x_1)]^{1/2}$$

$$\beta_{11}=1/\|\psi_1\|=1/[\varphi_1(x_1)]^{1/2}, \varphi_1^*(x)=\beta_{11}\varphi_1(x)$$

(b) for k=2, based on the results of step (a), compute:

$$(\varphi_2(x),\varphi_1^*(x))=\beta_{11}\varphi_2(x_1)$$

$$\|\psi_2\|^2=\varphi_2(x_2)-[\beta_{11}\varphi_2(x_1)]^2$$

thus, β_{12} and β_{22} can be computed as:

$$\beta_{22}=\frac{1}{\sqrt{\varphi_2(x_2)-[\beta_{11}\varphi_2(x_1)]^2}}$$

$$\beta_{12}=-\beta_{22}\beta_{11}^2\varphi_2(x_1)$$

$$\varphi_2^*(x)=\beta_{12}\varphi_1(x)+\beta_{22}\varphi_2(x)$$

(c) for 1~*k*-1, suppose all $\beta_{ji}(j=1,..,i,\ i\le k-1)$ were computed and available, the left work is to compute $\beta_{jk}, j=1,...,k$.

For $j<k$, we have $\varphi_j^*(x)=\sum_{i=1}^{j}\beta_{ij}\varphi_i(x)$

$$(\varphi_k(x),\varphi_j^*(x))=\sum_{i=1}^{j}\beta_{ij}(\varphi_k(x),\varphi_i(x))$$

$$=\sum_{i=1}^{j}\beta_{ij}\varphi_k(x_i), \text{ (j<k)}$$

$$\|\psi_k\|^2=\varphi_k(x_k)-\sum_{j=1}^{k-1}\left[\sum_{i=1}^{j}\beta_{ij}\varphi_k(x_i)\right]^2$$

$$\beta_{kk}=1/\|\psi_k\|$$

for $j=1,...,k-1$, we can obtain β_{jk} as follow:

$$\beta_{jk}=-\beta_{kk}\sum_{i=j}^{k-1}\left[\beta_{ji}\sum_{m=1}^{i}\beta_{mi}\varphi_k(x_m)\right], j=1,...,k-1$$

(d) Computing the decomposition coefficients $\{\alpha_j\}_1^n$.

For j=1,…, n, do

$$\alpha_j=\sum_{k=1}^{j}\beta_{kj}f_k$$

and $f(t)=\sum_{i=1}^{n}\alpha_i\varphi_i^*(t)$

Let us make a brief analysis about the above algorithm. Suppose *N* samples, the computation of $\beta_{kk}(k>1)$ requires O(N^2) product operations, and $\beta_{jk}, j=1,...,k$ requires O(N^2) product operations too. Hence, the total product operations requires O(N^3). In the above analysis, the computation load to evaluate $\varphi_k(x_j)(k=1,...,N, j=1,...,k)$ is not taken account. Besides, the total add operations requires O(N^3) too.

Ⅳ. Experimental Results

According to the decomposition algorithm in above section, we will decompose some speech signals in space $W_2^1[a,b]$. First, the coefficients $\beta_{jk}, j=1,...,k$ will be calculated. Let's take the interval [0,1] and equally divide it into *n* parts, thus a set $\{x_j\}_1^n$ was obtained, where $\{x_j\}_1^n \subset [0,1]$, and the value of *n* was set to 16,32,320,512,1000, respectively. Now, for each value of *n*, a group of the coefficients $\beta_{jk}, j=1,...,k$ were calculated as shown in Table 5.1. By observing Table 5.1, we find the except β_{11} ≈0.8, other $\beta_{kk}(k>1)$ were within the range 4.12~31.63, and the values of $\beta_{k-1,k}(k>1)$ were almost equal to $\beta_{kk}(k>1)$, but the

sign was opposite. The values of other β_{jk}, $j \le k-2$ were very small, less than 10^{-10}, which could be ignored without any significant influence.

Table 5.1 Samples of β_{jk}, $j = 1,...,k$

N node	β_{11}	$\beta_{kk}\,(k>1)$	$\beta_{k-1,k}$	$\beta_{jk}\,(j<k-1)$
8	0.91	2.95	-2.72	10^{-15}
16	0.89	4.12	-3.91	10^{-14}
32	0.88	5.71	-5.62	10^{-13}
320	0.87	17.91	-17.87	10^{-12}
512	0.87	22.64	-22.61	10^{-11}
1000	0.87	31.63	-31.61	10^{-11}

From above Table 5.1, we know that only β_{kk} and $\beta_{k-1,k}$ are important in practical computation. The following figure Fig. 5.1 showed the distribution of the values of β_{jk}, $j = 1,...,k$, where the node number n=32. The peaks of the comb shape are the values of non-zero β_{kk} and $\beta_{k-1,k}$. The program for computing the coefficients β and plotting the following figure Fig. 5.1 could be found in Appendix B and run with Mathematica 5.1.

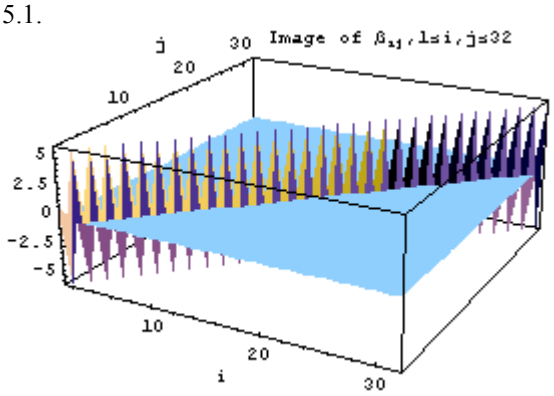


Fig. 5.1 Distribution of β_{ji}

In the computation process of β_{jk}, $j = 1,...,k$, only the node set $\{x_j\}_1^n$ and the RPK function set $\{\varphi_i(x)\}_1^j$ were applied, without any information of speech signal $f(x)$. So, these values can be applied to the processing of other speech frames and needn't to be re-computed.

Next, we will compute the decomposition coefficients α_i which used the sample values of speech signal $f(x)$. Since β_{ji} has localization property, the computation of α_i only requires $\beta_{ji}, \beta_{i-1,i}$ and $f(x_i)$, $f(x_{i-1})$. We set node number or frame length n=512, and computed the decomposition coefficients α_i of phoneme "i" and "sh" of Pinyin, respectively. The following Fig.5.2(a)~(b) showed the original waveform and the corresponding decomposition coefficients α_i of phoneme "i". The program for computing the coefficients α and plotting the following figure Fig. 5.2(a)~(b) could be found in Appendix A and run with Mathematica 5.1.

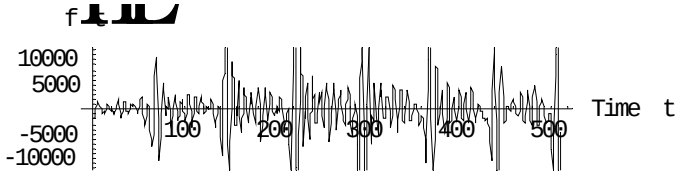


Fig. 5.2 (a) The original waveform of "i"

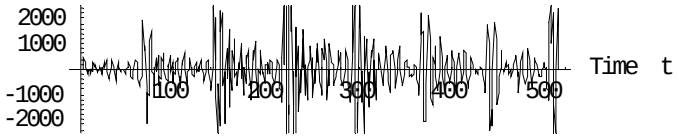


Fig. 5.2 (b) The coefficients α_i of "i"

From above Fig.5.2(a)~(b), we can find that the images of the original waveform and the coefficients α_i are very similar, and the image of α_i keeps the pitch features of the original waveform unchanged.

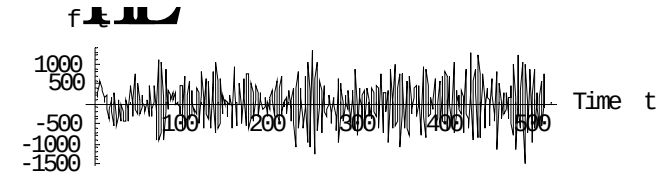


Fig. 5.3 (a) The original waveform of "sh"

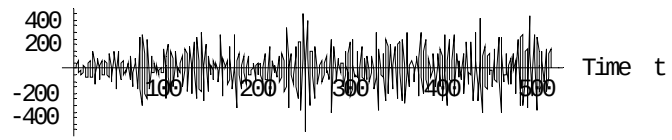


Fig. 5.3 (b) The coefficients α_i of "sh"

From above Fig.5.3(a)~(b), we can find that the images of the original waveform and the coefficients α_i are very similar, and the image of α_i keeps the noise-like features of the original waveform unchanged. Besides, the reconstructed speech waveform by Th.2.5 was identical to the original speech signal.

V. Conclusions

Speech signal orthogonal decomposition approach in the RPK space $W_2^1[a,b]$ was discussed and the realization algorithm was presented in this paper. In the RPK space $W_2^1[a,b]$, the speech signal based on some discrete samples was mapped into a continuous function, which is an exact analytical representation of the original speech signal. Such analytical representation is different from the canonical representation of speech signals by the series of triangle functions. Besides, the inner product computation in $W_2^1[a,b]$ was transformed into the

evaluation of some functions at some fixed points, which significantly improved the computation process.

ACKNOWLEDGMENT

The authors wish to thank Professor Cui Minggen. This work was supported in part by a grant from NSFC (No.60572125).

REFERENCES

[1] V.N. Vapnik, The Nature of Statistical Learning Theory, New York: Springer-Verlag, 1995

[2] Wahba, G. Reproducing Kernel Hilbert Spaces - Two Brief Reviews, in the Proceedings of the 13th IFAC Symposium on System Identification, 2003, 549-559.

[3] T. Kailath, An RKHS approach to Detection and estimation Problems - part v: Parameter Estimation, IEEE Trans. Information Theory, vol.IT-19, no.1, 1973, 29-37.

[4] Carl J. Nuzman and H. Vincent Poor, Reproducing Kernel Hilbert Space Methods for Wide-Sense Self-similar Processes, The Annals of Applied Probability, 2001, Vol.00, No.0, 1-21.

[5] F.M.Larkin, Optimal Approximation in Hilbert Space with Reproducing Kernel Function, Math. Comp., 1992, 22(4):911-921.

[6] Weinert, H.L., ed., Reproducing Kernel Hilbert Spaces: Application in Statistical Signal Processing, Hutchinson Ross, Stroudsburg, PA, 1982.

[7] Shantanu Chakrabatty, Yunbin Deng and Gert Cauwenberghs, Robust Speech Feature Extraction by Growth Transformation in Reproducing Kernel Hilbert Space, Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing (ICASSP'2004), Montreal Canada, May 17-21, 2004.

[8] A. Kocsor and L. Toth, Kernel-Based Feature Extraction with a Speech Technology Application, IEEE Trans. Signal Processing, Vol. 52, No. 8, Aug. 2004, 2250-2263.

[9] Kocsor, A. and Toth, L., Application of Kernel-Based Feature Space Transformations and Learning Methods to Phoneme Classification, Applied Intelligence, Vol. 21., No. 2., pp. 129-142, 2004.

[10] J. Shawe-Taylor and N. Cristianini, Kernel methods for pattern analysis. Cambridge University Press, 2004.

[11] Toth, L., Kocsor, A., Harmonic Alternatives to Sine-Wave Speech, Proc. Eurospeech, Geneva, 2003, 2073-2076.

[12] Toth, L., Kocsor, L., Gosztolya, G., Telephone Speech Recognition via the Combination of Knowledge Sources in a Segmental Speech Model, Acta Cybernetica, Vol 16, pp. 643-657, 2004

[13] Toth, L., Kocsor, A., Explicit Duration Modelling in HMM/ANN Hybrids, Matousek et al. (eds.): Proceedings of TSD 2005, LNAI 3658, pp. 310-317, Springer, 2005

[14] Banhalmi, A., Kovacs, K., Kocsor, A., Toth, L., Fundamental Frequency Estimation by Least-Squares Harmonic Model Fitting, Proceedings of EuroSpeech 2005, pp. 305-308.

[15] A. Kocsor, L. Toth, et al., A Comparative Study of Several Feature Transformation and Learning Methods for Phoneme Classification, Int. Journal of Speech Technology, vol.3, no.3/4, 2000, 263-276

[16] Cui Minggen, Wu Boying, Numerical Analysis of Reproducing Kernel Space, Science Press (China), 2004.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 349-352

A New Development Architecture for E-Commerce Platform

Longjun Huang [1] , Caiying Zhou [2], and Yuanwang Wei [3]
[1]Software of Software,JiangXi Normal University ,NanChang,China
Huanglong@ jxnu.edu.cn
[2]Science&Technology Division ,JiangXi University of Science and Technology ,GanZhou, China
Zhoucaiying_007@ mail.jnust.cn
[3]Department of Computer, JiaXing University, JiaXing, ZeJiang, China
yuanwang_wei@ 163.com

***Abstract*— E-commerce is a kind of commercial activity which adopts electronic form under the condition of open Internet. It has become an important life style. The diversity and variability of e-commerce activities means a great challenge to software development. How to design and develop a flexible and reusable e-commerce platform has become a direction that e-commerce industry is heading to. Based on the three-tier architecture, a new e-commerce platform development framework-seven layer architecture has been given in this paper. It describes the design ideas and the concrete realization of each layer. Practice shows that the development architecture proposed here can meet the needs of the diverse and fast-changing e-commerce business. Therefore, it's a feasible solution.**

***Index Terms*— E-commerce;Seven layer architecture; Page structure layer; Bookstore**

I. INTRODUCTION

Nowadays, to conduct transactions through e-commerce platform has taken a great role in modern society. E-commerce have entered into all aspects of society, it has a variety of requirements which are easy to change from users, demands different implementation approach, even different deployment platform. Traditional developing methods are not competent. Even if it is realized, the high degree of system coupling, is not easy to change and difficult to maintenance. There are still some other problems which make it far from being able to meet the needs of e-commerce field. It has proposed three-tier architecture which divided the entire business applications into three layer: presentation layer (UI), business logic layer (BLL), Data Access Layer (DAL). The three-tier architecture fully reflects the "high cohesion, low coupling" thinking and it solve many problems encountered in the process of complex system development [1]. Although this architecture improve the encapsulation and maintainability of the project, but the encapsulation still not go far enough, each layer still have some content which could be separated and the integrity of the system is inadequate too. Based on the above analysis, a new development architecture - seven layer architecture has been proposed to respond the e-commerce business features. This paper discusses the system framework, the main layer's design of this architecture, and gives the implementation of an online bookstore based on this architecture.

II. THREE TIER ARCHITECTURE[2][3]

Layered application designs are extremely popular because they increase application performance, scalability, flexibility, code reuse, and other benefits. In the classic three tier design, applications break down into three major areas of functionality: the data access layer, the business layer and the presentation layer. Inside each tier there may also exists a series of sub-layers that provide an even more granular breaking up the functional areas of the application.

Data access layer (DAL): the data access layer manages the physical storage and retrieval of data. In short words, it is to perform Select, Insert, Update, and Delete of the data table.

Business logic layer (BLL): As the essential part of the whole system, it maintains business rules and logic. The relevant design of business logic tier is associated with the unique logic of e-commerce, such as merchandise inquiries, making orders; adding merchandise to shopping carts etc. If it involves access to databases, then transfer the data access layer.

Presentation layer: the presentation layer houses the user interface and related presentation code. In this layer, ideal system status should not include the business logic. The logic codes in the presentation layer are only relating to the interface elements.

III. SEVEN LAYER ARCHITECTURE

A. Overview

1) Background

Based on three-tier architecture, we make the data access layer and the presentation layer into detailed layers. The DAL is divided into physical data layer, data access layer and entity layer. The UI is divided into page structure layer, logical control layer, and page presentation layer. This layered approach is an optimum approach for three-tier structure. By using this developing model, developers can only concerned with on one layer in the whole system, any layer's implementation can easily be replaced with a new realization, and the dependence

between layers is reduced. All this features will conduce to standardization and the reuse of each layer.

2）The Definition of Each Layer

Physical data layer: it includes business modeling and database design.

Data access layer: it includes the design and implementation of data access components.

Entity layer: the mapping between objects and data

Business logic layer (BLL): As the essential part of the whole system, it maintains business rules and logic based on user requirements, includes adding, deleting, updating, and selecting methods which are the guarantee of finishing the system.

Page structure layer: it includes page element such as HTML, but not the style which is described in page presentation layer.

Logical control layer: it control the page's content by dominating the page structure layer, it access the business logic layer.

Page presentation layer: this layer use div+css to meet the user's requirements of the page surface, it separate from page structure layer.

The overall structure and relations between each layer are showed at figure 1:

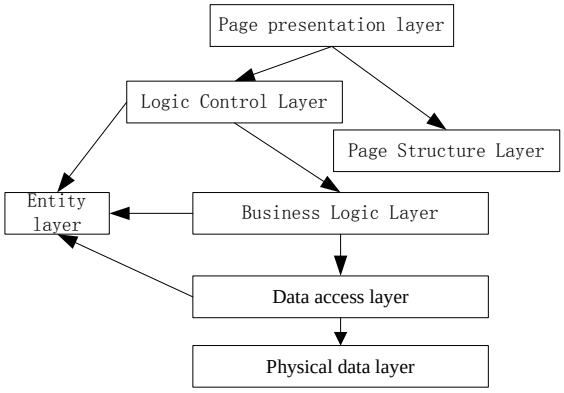


Figure 1. The Overall Structure of Seven Layer Architecture

3）Signification

By using the seven layer architecture, if errors occurred in the system, we can easily find the layer where the error occurred, and can quickly find the exact location of the error in the layer and then amend it, because the encapsulation degree is very high in our architecture. The flexibility of seven layer architecture is reflected in the reusability, scalability and maintainability.

B. Physics Data Layer

In this layer, it mainly includes the establishment of basic data tables, the realization of data integrity, the encapsulation of complex business logic. This layer is competed by systems analysts, database administrators, and experts in the field together.

C. Data Access Layer

A Data Access Layer (DAL) is a layer of a computer program which provides simplified access to data stored in persistent storage of such kind, such as an entity-relational database. For example, instead of using commands such as insert, delete, and update to access a specific table in a database, a class and a few stored procedures could be created in the database. The procedures would be called from a method inside the class, which would return an object containing the requested values. Or, the insert, delete and update commands could be executed within simple functions like register user or login user stored within the data access layer [4].

D. Entity Layer

Object-oriented thinking is now generally used for development, while the data is often stored in a relational database. The physical layer's main purpose is to resolve this contradiction, it maple relational date into objects. Often the entities are generated by code generator. It is a mistake that the entities are just the mapping tables. If the domain model is considered, the entities will different from tables and will be solved by mapping. As domain model, entities should consider more about domain itself.

E. Business Logic Layer

As the essential part of the whole system, business logic tier maintains business rules and logic. The relevant design of business logic tier is associated with the unique logic of e-commerce, such as merchandise inquiries, making orders; adding merchandise to shopping carts etc. If it involves access to databases, then transfer the data access layer.

F. Page Structure Layer

The traditional page structure and appearance are mixed together. If you want to modify the appearance of the page, the workload will become very large. This page is not in conformity with maintainability. So in the page structure layer we just define the page level, put all the required data elements on the page by tags and without considering the decoration of the page. Although the page appears very stiff and monotonous, however, this way makes it very easy to produce and change page's appearance by using style.

G. Logic Control Layer

Logic control layer is about the nexus. On one hand, it controls the page structure layer which aim is to collect data; on the other hand, it uses the business methods provided by BLL and makes the logic embodied in page structure layer, aims to load data.

H. Page Presentation Layer

The advantages of Div+css layout are being able to greatly improve the maintainability of the page by separating the page structure layer and the page presentation layer; this is the benefits of xhtml too. When the customer requirements for the appearance of the page change, you can easily make changes to the css part, thus changing the overall effect of the page, but the page itself need not to be changed.

IV. ADVANTAGES

Both seven-tier and three-tier architecture are based on hierarchical design ideas. They fully reflect the "high

cohesion, low coupling" thinking, solve many problems encountered in the process of complex system development and improve the encapsulation and maintainability of the project. Compared to three-tier architecture, seven–tier architecture divides the data access layer and the presentation layer into more detailed layers. In this architecture, the page presentation designer and the business designer are separated; the DBA and the programmer are separated too. By reducing the dependence between layers, it will conduce to standardization and the reuse of each layer and so improve the efficiency of system development and maintenance.

V. APPLICATION

An online bookstore is developed based on seven layer architecture mentioned above; it has following functions [5]:

Book Management: This module mainly includes book category management; book adding, modifying, deleting, inquiring; book inventory management, book comment browser as well as other management functions.

Order Management: This module mainly includes the purchase of books, order generation, automatic delivery and orders modification. Here the order generation is the kernel module of this system, which put the information acquisition and processed data into database after every order instantiation. It aims at tracking each transaction

Customer Management: It is the pivotal step in designing an online shopping website, and mainly used to manage the general information as well as the status of the vast number of registered members by putting their information into database. After registration, users can log in to shopping under their name and password on websites.

Forum Management: It is very important for a good e-commerce website to provide customers with effective ways of communication. This book store online system has designed a full-fledged forum for users to communicate.

Decision-making support: Most of e-commerce websites only provide online transactions functions. However, in this system, users are provided with a wealth of functions assisting decision-making, and much useful information useful to making decisions can be achieved through the analysis of many sales databases, such as the charts of sales of books, the shopping tendencies of specific customer base etc.

System Maintenance: This module is mainly used to do the maintenance of some general information of the users and manage users' privileges and database security in order to ensure the safety of the system by preventing unauthorized users' from causing damage to the program intentionally or unintentionally[6].

VI. THE IMPLEMENT OF EACH LAYER

A. *The Realization of Physical Data Layer*

Our system design a total of 21 tables, including user, books, sales, rights management etc, in addition to a large number of stored procedures, views, and so on.

B. *The Achievement of Data Access Layer*

This layer is aimed to encapsulate the access to the database. According to the interface-oriented programming ideas, we define the data access interface at data access layer, the upper layers program to interface and transparency to physical database. The project at this stage just uses SQL server2005, so it achieved data access interface class only for SQL server as showed at fig 2.

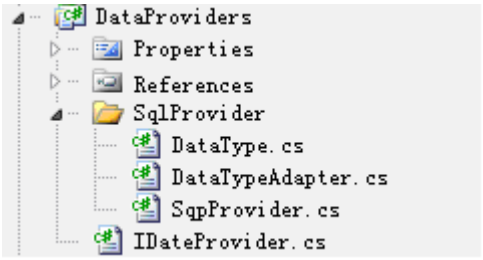


Figure 2. The design of Data Access Layer

Data providers which are encapsulated in this project are much simpler to native data providers-ADO.Net. A large number of database operations are encapsulated in the base class. The programmer just call method in this class, this practice reflects factory pattern of design pattern. Encapsulation not only reduces the number of codes, but also reduces the probability of making errors.

C. *The Implement of Entity Layer*

As the bridge between relational database and objects, entity plays as a data transmission carrier at each layer. In the entity's constructor function, we use PropertyInfo technology in reflection to construct entities. When constructing, it receive a parameter (DataTable), then traverse the columns in the table and the attributes in the entity ,if them equals, give the column's value to the attribute automatically, so the duplication of code is reduced. This model not only improves the working efficiency, but also reduces the chance of error. The following is the main code for base class[7].

```
public Base(System.Data.DataTable table ,int rowindex)
{
        // Getting the attributes of the object
    System.Reflection.PropertyInfo[] pinfo = this.GetType().GetProperties();
    //Looping the columns
    for (int i = 0; i < table.Columns.Count; i++)
    {
      // Finding the attribute which has same name and type with the columns
        for (int j = 0; j < pinfo.Length; j++)
        {
          if (table.Columns[i].ColumnName == pinfo[j].Name && table.Columns[i].DataType == pinfo[j].PropertyType )
            {
              if(Convert.IsDBNull (table.Rows[rowindex][i])==false )
              {
```

```
            //Giving the column's value to the attribute which has same name and type.
            pinfo[j].SetValue(this, table.Rows[rowindex][i], null);
        -----
    }
```

D. The Implement of BLL

Business Logic Layer is undoubtedly the core part in the system architecture. It mainly focuses on the system design such as the formulation of business rules, the implement of business flows and so on which are related to business requirement. In other words, it is concerned with the system's domain logic. In this system, business logic layer in every module is encapsulated in a subproject- Components. The web layer calls the approach in business logic layer to achieve the appropriate action including basic adding, modifying, deleting, and a variety of query operations.

E. The Achievement of User Interface Layer

In this project, the user interface layer is divided into page structure layer, page, page presentation layer and logical control layer. By the separation between presenting layer and structure layer, the project have scalability, maintainability. Page structure layer is described by html, logic control layer is realized by the asp.net file and the page presentation layer is achieved by CSS.

VII. CONCLUSION

E-commerce is booming and more attention will be paid to the response speed of website, the security of customers' data, the stability and cost of system operating by future e-commerce enterprise class. It is very important to choose a flexible developing architecture when set up e-commerce platform.

By applying the tiered technical design system, the seven layer architecture under discussion in this paper enjoys a rigid structure, clear logic and small coupling between layers .This architecture can well adapt to system maintenance, scalability. An online bookstore has been developed by using the seven layer architecture and works well. Practice has proved that it is a workable architecture.

ACKNOWLEDGMENT

We are grateful to Professor Huang Minghe for his help to our work. We especially want to express our gratitude to the authors of the referenced papers. In many respects, this paper represents the blending of insights gained from their research work.

REFERENCES

[1] ZhangYi. The Design of PetShop's system architecture [EB/OL]. http:// www. cnblogs.com/wayfarer/archive/2006 /04/14/375382. html,2009-02-05

[2] Microsoft. Deployment Patterns[EB/OL].http: //msdn.mic-rosoft.com/en-us/library/ms998478.aspx,2009-10-12

[3] Huang Longjun, Zhou Caiying,Dai Liping, Huang Minghe. Research and Implementation of E-commerce Platform Based on .NET Framework[C]. Proceedings of the 2009 International Symposium on Web Information Systems and Applications (WISA'09)Nanchang, P. R. China, May 22-24, 2009, pp. 112-115

[4] Jeffrey Richter. Applied Microsoft.NET Framework programming [M].World Book Publishing Company, 2002

[5] Li mengfei, Gao qijuan etc. The Design of E-commerce platform based on MVC Pattern. China CIO News[J]. 2008,10:66-67

[6] Kern A.Advanced Features for Enterprise-wide Role-based Access Control[C].18th Annual Computer Security Applications Conference,2002:333-342

[7] Solis,D. Illustrated C# 2008[M]. Turing,2008

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 353-358

Reproducing Kernel Functions Represented by Form of Polynomials

Sen Zhang, Lei Liu, and Luhong Diao
College of Applied Sciences, Beijing University of Technology, Beijing, China
Email: zhangsen@yahoo.com

***Abstract*—By re-defining the inner product of a reproducing kernel space, the reproducing kernel functions of that space can be represented by form of polynomials without changing any other conditions, and the higher order of the derivatives, the simpler of the reproducing kernel function expressions. Such expressions of reproducing kernel functions are the simplest from the computational point of view, resulting in speed and accuracy significant improvement in scientific and engineering applications. The performance of such reproducing kernel functions is shown to be very encouraging by experimental results.**

***Index Terms*—Reproducing Kernel Space; Reproducing Kernel Function; Hilbert Space**

I. INTRODUCTION

Bergman [1–7], J. Mercer [8], E. H. Moore and S. Bochner respectively proposed a special function in the 1920s in their different research fields, i.e., Bergman called $\tilde{E}(z,\overline{z},t)$ as the producing function of differential equations (now it is also called Bergman kernel function), J. Mercer named $K(x,y)$ as positive definite kernel, etc. In the 1950s, N. Aronszajn [9] summarized the previous related research results and used "reproducing kernel function" as the identical term for these different functions, so the foundations of reproducing kernel theory was set up. Since then, many researchers have contributed much work to the development and improvement of reproducing kernel theory.

In 1986, Cui [10] proved that $W_2^1[a,b]$ is a Hilbert space with reproducing kernel function, and exactly expressed the reproducing kernel function of $W_2^1[a,b]$ by finite terms. Hence, the application of reproducing kernel theory began in many areas. In recent years, many research reports showed that some problems could be solved in reproducing kernel space $W_2^m[a,b]$ because the reproducing kernel function of $W_2^m[a,b]$ could be exactly expressed by finite terms. For instance, the problem of infinite linear equations, the singular boundary problem, the period problem, the nonlinear problem and the inverse problem of wave equation could be solved in $W_2^m[a,b]$ with satisfied solutions [11-19].

However, the expression of the reproducing kernel function of $W_2^m[a,b]$ used in previous papers is too complicate, and the complexity of the expression of the reproducing kernel function of $W_2^m[a,b]$ will increase while *m* becomes larger, which will lead to some significant difficulties in computation. For example, the reproducing kernel function of $W_2^7[0,1]$ is a segmental function expressed by $x^k, e^{\lambda x}, \sin\alpha x, \cos\beta x$ and some fundamental operations (such as add, subtract, multiply and divide, etc), and could be full of more than 9 or 10 A4 pages if printed out. Moreover, the higher of differential orders, the more complicate of the reproducing kernel function. Therefore, the complicate reproducing kernel function could lead to some serious problems in computation, and some problems may become very difficult if the function in these problems with high smoothness degrees.

Based on the reproducing kernel space $W_2^m[a,b]$ established by Professor Cui Minggen and by re-defining the inner product, the reproducing kernel functions of $W_2^m[a,b]$ can be significantly simplified and expressed by polynomials without changing any other conditions. In this case, the reproducing kernel functions could be represented by piecewise polynomials, and the higher order of derivatives, the simpler of the reproducing kernel function expressions. Such expressions of reproducing kernel functions are the simplest from the computational viewpoint, the speed and accuracy could be significantly improved in scientific and engineering applications. The performance of such reproducing kernel functions is shown to be very encouraging by experimental results.

II. RPKS $W_2^m[a,b]$

The function space $W_2^m[a,b]$ is defined as follows:

$$W_2^m[a,b]=\{f(x)\mid f^{(m-1)}(x) \text{ is absolutely continuous},\ f^{(m)}(x)\in L^2[a,b],\ x\in[a,b]\} \quad (2.1)$$

The inner product and the norm in the function space $W_2^m[a,b]$ are defined as follows respectively:

for any functions $f(x), g(x)\in W_2^m[a,b]$,

$$<f,g>=\sum_{i=0}^{m-1} f^{(i)}(a)g^{(i)}(a)+\int_a^b f^{(m)}(x)g^{(m)}(x)dx, \quad (2.2)$$

$$\|f\|=\sqrt{<f,f>} \quad (2.3)$$

It is easy to prove that $W_2^m[a,b]$ is an inner space with the definitions of (2.2).

Theorem 2.1 Function space $W_2^m[a,b]$ is a Hilbert Space.

Proof: Suppose $f_n(x)$ $(n=1,2,...)$ is a Cauchy sequence in $W_2^m[a,b]$, i.e., if $n\to\infty$, then

$$\|f_{n+p}-f_n\|^2=\sum_{i=0}^{m-1}[f_{n+p}^{(i)}(a)-f_n^{(i)}(a)]^2+\int_a^b[f_{n+p}^{(m)}(x)-f_n^{(m)}(x)]^2dx\to 0$$

.

Therefore, we have $f_{n+p}^{(i)}(a)-f_n^{(i)}(a)\to 0$,

$i=0,1,\cdots,m-1$, (2.4)

and $$\int_a^b[f_{n+p}^{(m)}(x)-f_n^{(m)}(x)]^2dx\to 0. \quad (2.5)$$

which indicates that for any i ($0\le i\le m-1$), the sequence $f_n^{(i)}(a)$ $(n=1,2,...)$ is a Cauchy sequence and $f_n^{(m)}(x)$ $(n=1,2,...)$ is a Cauchy sequence in space $L^2[a,b]$. So, there exist unique real number λ_i $(i=0,1,...,m-1)$ and unique function $h(x)\in L^2[a,b]$, satisfy the following:

$$\lim_{n\to\infty}f_n^{(i)}(a)=\lambda_i \ (0\le i\le m-1)$$

and $$\lim_{n\to\infty}\int_a^b[f_n^{(m)}-h(x)]^2dx=0.$$

Suppose

$$g(x)=\sum_{k=0}^{m-1}\frac{\lambda_k}{k!}(x-a)^k+\overbrace{\int_a^x\int_a^x\cdots\int_a^x}^{m}h(x)(dx)^m, \quad (2.6)$$

since $h(x)\in L^2[a,b]$, hence $g^{(m-1)}(x)=\lambda_{m-1}+\int_a^x h(x)dx$ is absolutely continuous in $[a,b]$, and $g^{(m)}(x)=h(x)$ is true almost everywhere in $[a,b]$. So, $g(x)\in W_2^m[a,b]$ and $g^{(i)}(a)=\lambda_i$ $(0\le i\le m-1)$. Moreover, we have:

$$\|f_n(x)-g(x)\|^2=\sum_{i=0}^{m-1}[f_n^{(i)}(a)-g^{(i)}(a)]^2+\int_a^b[f_n^{(m)}(x)-g^{(m)}(x)]^2dx$$

$$=\sum_{i=0}^{m-1}[f_n^{(i)}(a)-\lambda_i]^2+\int_a^b[f_n^{(m)}(x)-h(x)]^2dx\to 0.$$

Hence, function space $W_2^m[a,b]$ is a Hilbert Space. □

Lemma 2.2 $W_2^m[a,b]$ is a reproducing kernel space if and only if for any $x\in[a,b]$, $I:f\to f(x)$ is a bounded functional in $W_2^m[a,b]$ [9].

Theorem 2.3 Function space $W_2^m[a,b]$ is a reproducing kernel space.

Proof: In fact, suppose $x\in[a,b]$ and $f(x)\in W_2^m[a,b]$, we have

$$f^{(m-1)}(x)=f^{(m-1)}(a)+\int_a^x f^{(m)}(x)dx,$$

and

$$|f^{(m-1)}(x)|\le|f^{(m-1)}(a)|+\int_a^x|f^{(m)}(x)|dx\le|f^{(m-1)}(a)|+\int_a^b|f^{(m)}(x)|dx,$$

where

$$\int_a^b|f^{(m)}(x)|dx\le\sqrt{(b-a)\int_a^b|f^{(m)}(x)|^2dx}=\tilde{M}_1\sqrt{\int_a^b|f^{(m)}(x)|^2dx}$$

$$\le\tilde{M}_1\sqrt{\sum_{i=0}^{m-1}[f^{(i)}(a)]^2+\int_a^b|f^{(m)}(x)|^2dx}=\tilde{M}_1\|f\|_{W_2^m}.$$

Note that for any i ($0\le i\le m-1$), we have

$$|f^{(i)}(a)|=\sqrt{[f^{(i)}(a)]^2}\le\sqrt{\sum_{i=0}^{m-1}[f^{(i)}(a)]^2+\int_a^b[f^{(m)}(x)]^2dx}=\|f\|_{W_2^m},$$

(2.8)

therefore $|f^{(m-1)}(x)|\le M_1\|f\|_{W_2^m}$. (2.9)

Note

$$|f^{(m-2)}(x)|\le|f^{(m-2)}(a)|+\int_a^x|f^{(m-1)}(x)|dx\le|f^{(m-2)}(a)|+\int_a^b|f^{(m-1)}(x)|dx$$ from (2.8) and (2.9), we have

$$|f^{(m-2)}(x)|\le\|f\|_{W_2^m}+(b-a)M_1\|f\|_{W_2^m}=M_2\|f\|_{W_2^m}$$

(2.10)

Similarly we have

$$|I(f)|=|f(x)|\le M_m\|f\|_{W_2^m} \quad (2.11)$$

So, I is bounded functional in $W_2^m[a,b]$ and $W_2^m[a,b]$ is a reproducing kernel space. □

Now, let's find out the expression form of the reproducing kernel function $R_m(x,y)$ in $W_2^m[a,b]$.

Suppose $R_m(x,y)$ is the reproducing kernel function of $W_2^m[a,b]$, then for any fixed $y\in[a,b]$ and any $f(x)\in W_2^m[a,b]$, $R_m(x,y)$ must satisfy the following:

$$< f(x), R_m(x,y) >= f(y) \qquad (2.12)$$

Based on (2.2), we have:

$$< f(x), R_m(x,y) >= \sum_{i=0}^{m-1} f^{(i)}(a) \frac{\partial^i R_m(a,y)}{\partial x^i} + \int_a^b f^{(m)}(x) \frac{\partial^m R_m(x,y)}{\partial x^m} dx$$

and

$$\int_a^b f^{(m)}(x) \frac{\partial^m R_m(x,y)}{\partial x^m} dx$$

$$= \sum_{i=0}^{m-1} (-1)^i f^{(m-i-1)}(x) \frac{\partial^{m+i} R_m(x,y)}{\partial x^{m+i}} \Big|_{x=a}^{b} + (-1)^m \int_a^b f(x) \frac{\partial^{2m} R_m(x,y)}{\partial x^{2m}} dx.$$

by variable substitution, we have

$$\sum_{i=0}^{m-1} (-1)^i f^{(m-i-1)}(x) \frac{\partial^{m+i} R_m(x,y)}{\partial x^{m+i}} = \sum_{i=0}^{m-1} (-1)^{m-i-1} f^{(i)}(x) \frac{\partial^{2m-i-1} R_m(x,y)}{\partial x^{2m-i-1}}$$

Moreover,

$$< f(x), R_m(x,y) >= \sum_{i=0}^{m-1} f^{(i)}(a) [\frac{\partial^i R_m(a,y)}{\partial x^i} - (-1)^{m-i-1} \frac{\partial^{2m-i-1} R_m(a,y)}{\partial x^{2m-i-1}}]$$

$$+ \sum_{i=0}^{m-1} (-1)^{m-i-1} f^{(i)}(b) \frac{\partial^{2m-i-1} R_m(b,y)}{\partial x^{2m-i-1}} + (-1)^m \int_a^b f(x) \frac{\partial^{2m} R_m(x,y)}{\partial x^{2m}} dx$$

.

Therefore, $R_m(x,y)$ is the solution of the following generalized differential equation:

$$\begin{cases} (-1)^m \dfrac{\partial^{2m} R_m(x,y)}{\partial x^{2m}} = \delta(x-y), \\ \dfrac{\partial^i R_m(a,y)}{\partial x^i} - (-1)^{m-i-1} \dfrac{\partial^{2m-i-1} R_m(a,y)}{\partial x^{2m-i-1}} = 0, \ i = 0,1,\cdots,m-1, \\ \dfrac{\partial^{2m-i-1} R_m(b,y)}{\partial x^{2m-i-1}} = 0, \ i = 0,1,\cdots,m-1. \end{cases}$$

(2.13)

While $x \neq y$, it is easy to know that $R_m(x,y)$ is the solution of the following constant linear homogeneous differential equation with $2m$ orders, i.e.,

$$(-1)^m \frac{\partial^{2m} R_m(x,y)}{\partial x^{2m}} = 0, \qquad (2.14)$$

with the boundary conditions:

$$\begin{cases} \dfrac{\partial^i R_m(a,y)}{\partial x^i} - (-1)^{m-i-1} \dfrac{\partial^{2m-i-1} R_m(a,y)}{\partial x^{2m-i-1}} = 0, \ i = 0,1,\cdots,m-1, \\ \dfrac{\partial^{2m-i-1} R_m(b,y)}{\partial x^{2m-i-1}} = 0, \ i = 0,1,\cdots,m-1. \end{cases}$$

(2.15)

We know that equation (2.14) has characteristic equation $\lambda^{2m} = 0$, and the eigenvalue $\lambda = 0$ is a root whose multiplicity is $2m$. Therefore, the general solution of equation (2.14) is as follows:

$$R_m(x,y) = \begin{cases} lR_m(x,y) = \sum_{i=1}^{2m} c_i(y) x^{i-1}, & x < y, \\ rR_m(x,y) = \sum_{i=1}^{2m} d_i(y) x^{i-1}, & x > y. \end{cases}$$

(2.16)

Now we are ready to calculate the coefficients $c_i(y)$ and $d_i(y)$, $i = 1,\cdots,2m$.

Since $(-1)^m \dfrac{\partial^{2m} R_m(x,y)}{\partial x^{2m}} = \delta(x-y)$.

Then we have :

$$\frac{\partial^i lR_m(y,y)}{\partial x^i} = \frac{\partial^i rR_m(y,y)}{\partial x^i} \quad i = 0,1,\cdots,2m-2 \quad ,$$

(2.17)

and

$$(-1)^m (\frac{\partial^{2m-1} rR_m(y+,y)}{\partial x^{2m-1}} - \frac{\partial^{2m-1} lR_m(y-,y)}{\partial x^{2m-1}}) = 1$$

(2.18)

The above equations in (2.17) and (2.18) provided $2m$ conditions for solving the coefficients $c_i(y)$ and $d_i(y)$ ($i = 1,\cdots,2m$) in equation (2.16). Note that equation (2.15) provided $2m$ boundary conditions, so we have $4m$ equations, i.e., (2.15), (2.17) and (2.18). It is easy to know these $4m$ equations are linear equations with the variables $c_i(y)$ and $d_i(y)$, and the $c_i(y)$ and $d_i(y)$ could be calculated out by many methods。As long as the coefficients $c_i(y)$ and $d_i(y)$ are known, the exact expression of the producing kernel function $R_m(x,y)$ of $W_2^m[a,b]$ could be calculated out from equation (2.16). The expression of $R_m(x,y)$ is a piecewise polynomial with $2m-1$ degrees.

Note: if the functions in space $W_2^m[a,b]$ require more special boundary conditions, e.g., the boundary conditions of the second order differential equations as follows:

$u(a) = \alpha, u(b) = \beta$, or $u(a) = \alpha, u'(a) = \beta$;

or the linear boundary conditions:

$a_1 u(a) + b_1 u'(a) = \alpha$, $a_2 u(b) + b_2 u'(b) = \beta$;

or the periodic linear boundary conditions:

$u(a) = u(b)$, $u'(a) = u'(b)$

These different kinds of boundary conditions could be contained in space $W_2^m[a,b]$ after homogenization, i.e., we can find the reproducing kernel function $R_m(x,y)$ which satisfies these boundary conditions. In all, the reproducing kernel space $W_2^m[a,b]$ has the simplest reproducing kernel function $R_m(x,y)$ represented by polynomials and could be applied in many areas.

III. RPK FUNCTIONS OF $W_2^m[0,1]$

We are ready to present some expressions of reproducing kernel function in $W_2^m[0,1]$ by using the approaches proposed in the above sections.

A. RPK function $R_1(x,y)$ in $W_2^1[0,1]$

$$R_1(x,y)=\begin{cases}1+x, & x\le y,\\ 1+y, & x>y,\end{cases}\quad x,y\in[0,1]$$

The 3-d and 2-d images of the reproducing kernel function $R_1(x,y)$ are shown in the following Fig. 3.1 and Fig. 3.2 respectively. In Fig. 3.1, $x,y\in[0,1]$, and Fig. 3.2, $x\in[0,1]$, y takes values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.9, respectively.

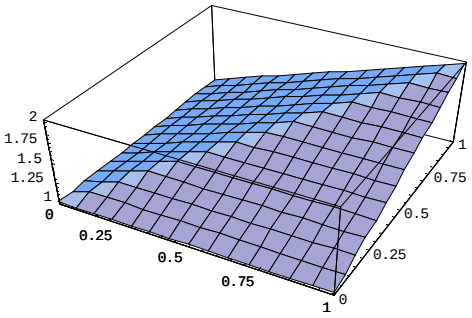


Fig. 3.1 3-d image of $R_1(x,y)$

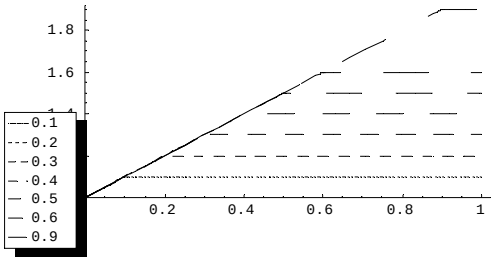


Fig. 3.2 2-d image of $R_1(x,y)$

B. RPK function $R_2(x,y)$ in $W_2^2[0,1]$

$$R_2(x,y)=\begin{cases}1-\dfrac{x^3}{6}+\dfrac{1}{2}xy(2+x), & x\le y,\\ 1-\dfrac{y^3}{6}+\dfrac{1}{2}xy(2+y), & x>y,\end{cases}\quad x,y\in[0,1].$$

The 3-d and 2-d images of the reproducing kernel function $R_2(x,y)$ are shown in the following Fig. 3.3 and Fig. 3.4 respectively. In Fig. 3.3, $x,y\in[0,1]$, and Fig. 3.4, $x\in[0,1]$, y takes values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.9, respectively.

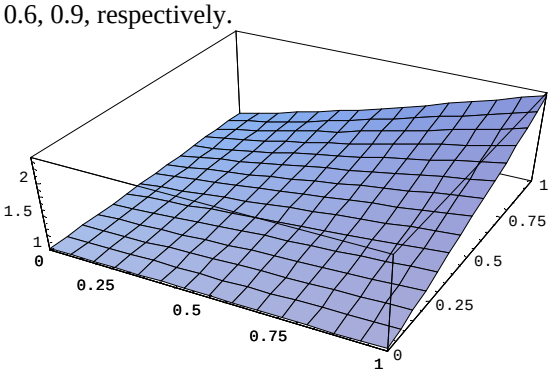


Fig. 3.3 3-d image of $R_2(x,y)$

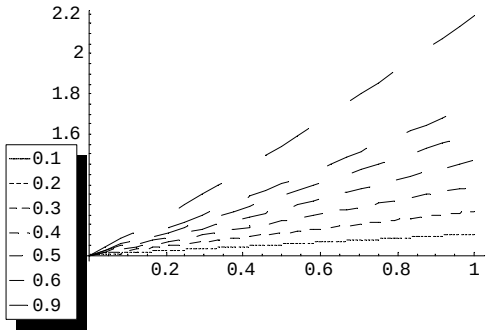


Fig. 3.4 2-d image of $R_2(x,y)$

C. RPK function $R_3(x,y)$ in $W_2^3[0,1]$

$$R_3(x,y)=\begin{cases}1+\dfrac{x^5}{120}+\dfrac{1}{12}x^2y^2(3+x)+xy(1-\dfrac{x^4}{24}), & x\le y,\\ 1+\dfrac{y^5}{120}+\dfrac{1}{12}x^2y^2(3+y)+xy(1-\dfrac{y^4}{24}), & x>y,\end{cases}\quad x,y\in[0,1]$$

D. RPK function $R_7(x,y)$ in $W_2^7[0,1]$

$$R_7(x,y)=\begin{cases}\dfrac{1}{6227020800}(6227020800+x^{13}+13x(479001600-x^{11})y\\ +78x^2(19958400+x^9)y^2+286x^3(604800-x^7)y^3\\ +715x^4(15120+x^5)y^4+1287x^5(336-x^3)y^5+1716x^6(7+x)y^6), & x\le y\\ 1+\dfrac{1}{6227020800}(y(y^{12}+1716x^6y^5(7+y)+1287x^5y^4(336-y^3)\\ +715x^4y^3(15120+y^5)+286x^3y^2(604800-y^7)\\ +78x^2y(19958400+y^9)+13x(479001600-y^{11}))), & x>y\end{cases}\quad x,y\in[0,1].$$

E. Numerical Examples

Some kinds of differential and integral equations with boundary or initial conditions have been studied and solved in literature [11-19] by using the properties of reproducing kernel functions, and the approximate solutions u_n (where n indicates the number of computing nodes) have been obtained. To test the performance of the reproducing kernel functions

presented in this paper, let's consider the following differential equation with boundary conditions:

$$\begin{cases} u''+u'+u=f(x),\ 0\le x\le 1 \\ u(0)=0, u'(0)=0. \end{cases} \qquad (4.1)$$

Let u_n be the approximate solution, u the exact solution, and $\varepsilon_n = \underset{x\in[0,1]}{Max}\{|u_n - u|\}$ indicate the absolute error. To solve the equation (4.1), a program was developed by using Mathematica 5.1 and run on a PC with CPU 2.0 GHz and RAM 512Mb. The experimental results are as follows:

（1）take the reproducing kernel function proposed in [11-19] and let $n=100$, set $\varepsilon_n \le 0.00005$, the running time is 759.547s.

（2）take the reproducing kernel function proposed in this paper and let $n=100$, set $\varepsilon_n \le 0.000014$, the running time is 7.204s.

（3）take the reproducing kernel function proposed in this paper and let $n=400$, set $\varepsilon_n \le 8\times 10^{-7}$, the running time is 148.61s.

The experimental results showed that the reproducing kernel function proposed in this paper could improve the computational speed significantly (more than 100 times). Moreover, we have proved that if $n\to\infty$, then $\varepsilon_n \to 0$. Thus, the computational accuracy could be improved too.

Ⅳ. Conclusions

It is well known that the Hilbert Space theory is the foundation of modern mathematics. The reproducing kernel space $W_2^m[a,b]$ is just a Hilbert Space with some special properties. So, $W_2^m[a,b]$ can inherit all the properties of the Hilbert Space and possess some special and better properties, which could make some problems be solved easier. For example, $L^2[a,b]$ is a complete Hilbert space . Many problems studied in $L^2[a,b]$ requires large amount of integral computations, and such computations may be very difficult in some cases. Thus, the numerical integrals have to be calculated in the cost of losing some accuracy. However, the properties of the producing kernel space $W_2^m[a,b]$ （see （2.12））require no more integral computation for some functions, instead of computing some values of a function at some nodes. This simplification of integral computation not only improves the computational speed, but also improves the computational accuracy.

Since N. Aronszajn put forward the reproducing kernel space theory in 1950s, many researchers have done much works in this field. Especially in recent 20 years, more and more experts have seen the advantages of reproducing kernel space. And that the reproducing kernel space $W_2^m[a,b]$ inherit all the properties of the Hilbert Space and possess some special and better properties, its widely application and better vision could be expected.

Acknowledgment

The authors wish to thank Professor Cui Minggen. This work was supported in part by a grant from NSFC (No.60572125).

References

[1] S.Bergman. Uber die entwicklung der harmonischen funcktionen der Ebene unddes raumes nach orthogonal funktionen. Math. Ann. 1922, 86:238-271

[2] S.Bergman.über Kurvenintegale von Funktionen zweier komplexen Veranderlichen, die Differentialgleichungen $\Delta V + V = 0$ befriedigen.Math.Z.32,1930:386-406

[3] S.Bergman.Über ein Verfahren zur Konstruktion der NaherungslÖSu-ngslÖsungen der Gleichung $\Delta u + \tau^2 u = 0$. Prikl. Mat. Meh. 1936:97-107

[4] S.Bergman.Zur Theorie der Funktion, die eine linear partielle Differentialgleichung berriedigen. Mat. Sb. 44, 1937: 1169-1198.

[5] S.Bergman.Zur Theorie der Funktion,die eine linear partielle Differentialgleichung. berriedigen. Soviet. Math. Dokl. 15, 1937:227-230.

[6] S.Bergman.Sur un lien entre lal théorie des equations aux derives Partielles elliptiques et celle des functions d'une variable complexe.C.R.Acal.Sci., Paris.205,1937:1198-1200,1360-1362.

[7] S.Bergman.The approximation of functions satisfying a linear partial differential equation.Duke Math.J., 1940,6:537-561.

[8] J.Mercer.Function of Positive and Negative Type and Their Connection with The Theory of Integral Equation.Philos. Trans.Roy. Soe.London Ser.A1909, 209:415-446.

[9] N.Aronszajn.Theory of reproducing kernels. Trans. Amer. Math. soc., 68, 1950:337-404.

[10] Minggen Cui, Zhongxing Deng, The optimal approximation operator in space $W_2^1[a,b]$, Journal of Computational Mathematics, 8:2(1986)209-216 (in Chinese)

[11] Lin Yingzhen, Cui Minggen, ZhengYi, Representation of the exact solution for infinite system of linear equation, Applied Mathematics and Computation, 168(2005)636-650.

[12] Chun-li Li, Minggen Cui, How to solve the equation AuBu+Cu=f, Applied Mathematics and Computation133 (2002) 643-653.

[13] Chun-li Li, Minggen Cui, The exact solution for a class nonlinear operator equations in the reproducing kernel space, Applied Mathematics and Computation 143(2003)393-399

[14] Fazhan Geng ,Minggen Cui, Solving singular nonlinear second-order periodic boundary value problems in the reproducing kernel space， Applied Mathematics and Computation, 192 (2007) 389-398.

[15] Minggen Cui, Fazhan Geng, Solving singular two-point boundary value problem in reproducing kernel space, Journal of Computational and Applied Mathematics, 205(2007)6-15

[16] Huanmin Yao, Minggen Cui, A new algorithm for a class of singular boundary value problems, Applied Mathematics and Computation, 186(2007)1183-1191.
[17] Lihong Yang, Minggen Cui, New algorithm for a class of nonlinear intergro-differential equations in the reproducing kernel space, Applied Mathematics and Computation 174(2006)942-960.
[18] Yunhui Li, Fazhan Geng, Minggen Cui, The analytical solution of singular linear periodic boundary value problem, Applied Mathematical Science, 1(2)(2007)77-87.
[19] Minggen Cui, Yingzhen Lin, A new method of solving the coefficient inverse problem of differential equation, Science in China Series A-Mathematics 4(2007).
[20] Minggen Cui and Boying wu, Numerical Analysis in Reproducing Kernel Space, Science Press (China), Beijing, 2004 (in Chinese)

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 359-364

A Distributed P2P Server System for Paper Sharing

Pingjian Zhang[1], and Juanjuan Zhao[1,2]
[1] School of Software Engineering, South China University of Technology, Guangzhou, China
Email: pjzhang@scut.edu.cn
[2] Nanchang Army College, Nanchang, China
Email: lzfjj@163.com

***Abstract*—P2P file sharing system is one of the hot research topics. However, most of such systems do not support auto extraction of metadata and provide only searches via resource titles. Combining the Chord algorithm and the SHA algorithm, this paper proposed a new P2P search model based on Distributed Hash Table (DHT). The procedures of establishing such P2P networks are presented. The design and implementation of the P2P Paper Sharing System is discussed. Experimental results show that the system achieves design goals.**

***Index Terms*—P2P Search Model, P2P Server System, Metadata Extraction, P2P Paper Sharing System**

I. INTRODUCTION

With the rapid development of information technology and the increase of the total volume of information, knowledge has played an essential role in the modern society. Knowledge management and sharing has become a hot research topic. As one of the most valuable knowledge resources, sharing of academic papers has important practical value, say, among research groups or developing teams. Being a resource sharing technology that has found a wide range of applications, the P2P networking [1-3] provides a convenient means of realization for sharing academic papers freely, which is a nice supplement to those commercial digital libraries. So far, many P2P resource sharing projects have been carried out, for example, Napster ([4]), Gnutella ([5]), KaZaa ([6]), Pastry ([7]), Maze ([8]) and Granary ([9]), to name just a few.

Although the P2P technology has obtained high-speed development in recent years, there are still some key issues to be solved ([10-15]). For example, the bandwidth occupation rate is high, the network expansibility is poor, and the resources usage is low, etc. The main reason lies in that resources in the P2P networks are of great dispersion, and nodes are free to join or exit, lacking unified and efficient management. How to effectively and reliably search resources in P2P networks becomes a challenging problem. In addition, most of the existing P2P resource sharing systems supports only searches by resource name, since resources are stored and manipulated in the whole document.

This paper presents a new P2P search model based on Distributed Hash Table (DHT), which consists of a layered server system. Ordinary server nodes (SN) form a chord ring, and a segment of SN of the ring is further managed by a super peep node (SPN), which is monitored by the coordinate node (CN). This layered model solves the problem of low search efficiency in the P2P network. To address the issue of metadata support when searching resources, the Chinese word segmentation module and the metadata auto extraction module are introduced. Experiments demonstrate that the new P2P paper sharing system works as desired.

The paper is organized as follows: in section 2, a new P2P search model is proposed along with the procedures to establish the P2P network. The design and implementation of the new P2P paper sharing system is discussed in section 3. Section 4 contains some final remark and conclusions.

II. A NEW P2P SEARCH MODEL

A. The P2P search model

The P2P search model consists of the following kinds of nodes:

- CN (coordinate node), that coordinates the resources managed by the SPN (super peep node) to prevent duplicate copies of resources.
- SPN, that is responsible for resources fetching, storage in FS (file server), and manage the join and exit of a group of SN (server node). SPN also reports routing information of SN to CN and broadcast routing information that it received from CN to its underlying SN.
- SN, that stores routing information about the whole P2P network and part of resource metadata in the (key, value) pairs. The SN form a chord ring.
- FS, that stores all the resources in the P2P network.

The topology for such a model is depicted in Fig. 1 below.

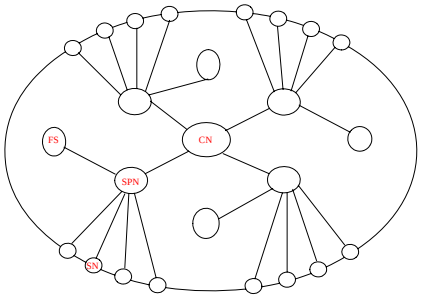


Figure 1 The P2P search model

The project is supported by the Guangdong Bureau of Science and Technology under the grant No. 2006B80107001.

B. The P2P networking strategy

The establishment of the P2P network consists of two steps:

1) Set-up of the SPN network

First of all, the CN starts up and initializes. Then, SPN registers to CN which generates a unique identifier and return it to the newly registered SPN. After receiving the registration response, SPN sends an acknowledgement to CN which will then notify all other SPN of the registration information. When the number of registered SPN arrive a certain threshold (which can be configured), the CN broadcasts a "finish" message to all SPN. The SPN then requests CN of routing information of all SPN and the CN responds with the required information. The process is described in Fig. 2 below.

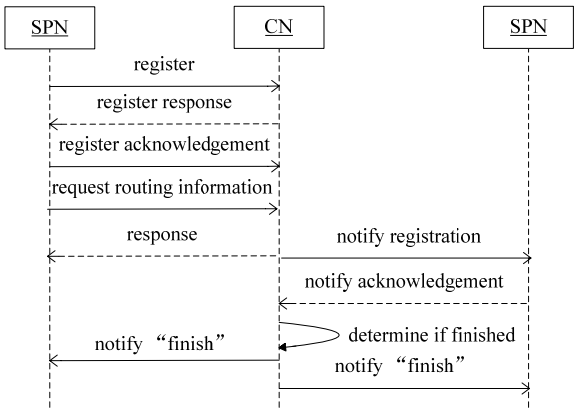


Figure 2 Sequence diagram of set up of SPN network

2) Set-up of the SN network

Only after SPN network has been set up can the SN network begin construction. At this stage, SN registers to SPN, which generates and returns a unique identifier by hashing the ip address of the SN via the SHA-1 algorithm. After receiving the registration response, SN sends an acknowledgement to SPN which will broadcast the routing information of the new SN to other SPN and then notify all other SN under its management of the registration information of the new SN. The new SN, in turn, will request its SPN of the global routing table.

When the number of registered SN arrive a certain threshold (which can be configured), the SPN broadcasts a "finish" message to all SPN including itself. When the number of "finish" message a SPN received equals the number of all SPN. The SPN then broadcasts the "whole network finish" message to all SN under its management. The process is described in Fig. 3 below.

3) Storage of resources

When SPN receives resources uploaded by clients, it stores it in the corresponding FS, extracts metadata and hashes them using SHA-1. Each (key, value) pair is sent to the successive SN nearest to the key in the chord ring, together with the information about the FS.

4) Search of resources

Upon the request of a resource, the SN extracts and hashes the metadata to obtain the key. Then, it queries the successive SN nearest to the key in the chord ring. The destination SN returns information about the FS that stores the desired resource. The SN fetches the resource from the FS and returns to the client.

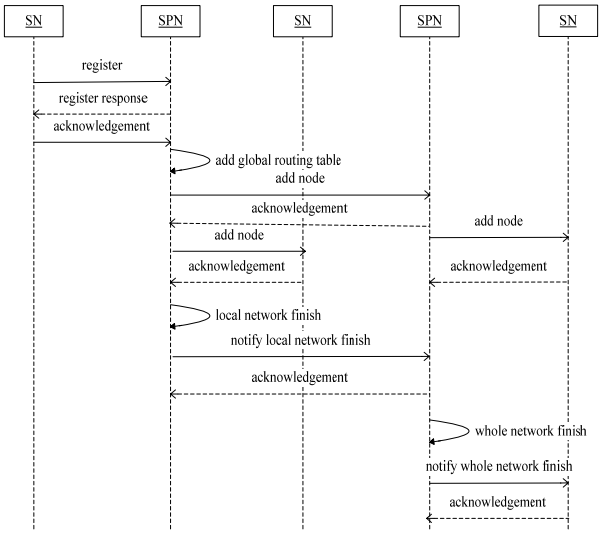


Figure 3 Sequence diagram of set up of SN network

5) Join of a new SN

After the whole network is constructed, if a new SN needs to join the network, it proceeds quite the same way as in the steps of set-up of the SN network.

6) Exit of an SN

When an SN wants to exit, it sends an "exit" message to its SPN which forwards the message to other SPN and SN under its management. Then, the successive SN to the exiting SN pulls data from the exiting SN and stores them. Upon receiving the "transfer finish" message, the exiting SN deletes data and exits, as shown in Fig. 4.

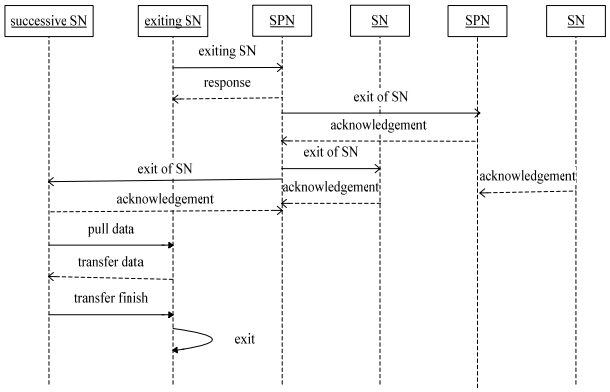


Figure 4 Sequence diagram of exit of an SN

III. Design and Implementation of the P2P Paper Sharing System

The system is designed to be a paper sharing system among research and development groups. Group members can upload valuable papers to share with others, and can search and download papers wanted.

The system consists of two subsystems: the P2P server system that falls into the P2P search model mentioned above, and the client end.

A. System analysis

The system is divided into 4 modules: the CN module, the SPN module, the SN module, and the client module, whose functionalities are list as follows.

The CN module will

1) handle registration of SPN and allocate a unique identifier for each SPN.

2) broadcast registration of new SPN to other SPN.

3) determine if the construction of SPN network is done.

4) assign SPN to clients.

The SPN module will

1) handle registration of SN.

2) broadcast routing information to other SPN and SN under its management.

3) store the global routing table.

4) process Chinese word segmentation, inverse document indexing and metadata hashing.

5) response to client query.

6) handle the exit of SN.

The SN module will

1) handle registration to SPN.

2) handle exit of SN.

3) store paper and metadata information.

4) synchronize with other SN.

The Client module will

1) extract metadata of papers and upload to SN together with the paper.

2) act as interface for searching papers via paper title and other metadata information.

B. Architecture of the system

Following the system requirement analysis, the system architecture consists of 4 subsystems and is interacts as in Fig. 5 below, while the deployment diagram shown in Fig.6.

C. Design and Implementation of the system

This subsection will highlight design and implementation issues of some key components.

1) The data dictionary class

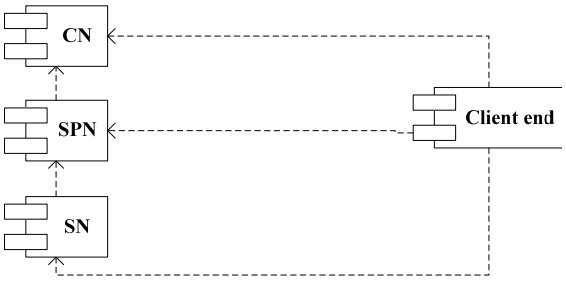


Figure 5 Component diagram of the P2P paper sharing system

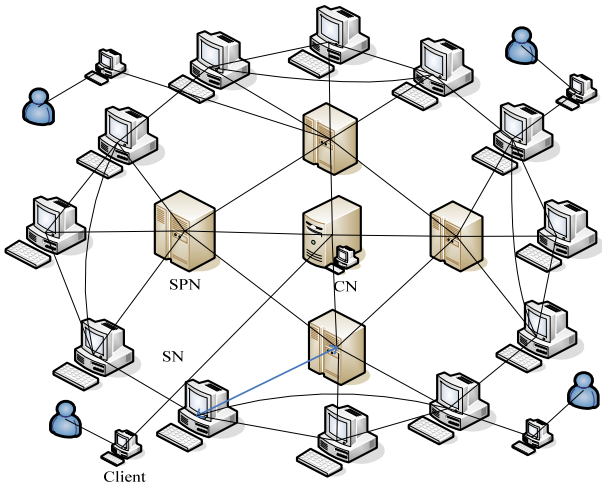


Figure 6 Component diagram of the P2P paper sharing system

Data dictionary is fundamental for the task of Chinese word segmentation. The class diagram is given in Fig. 7.

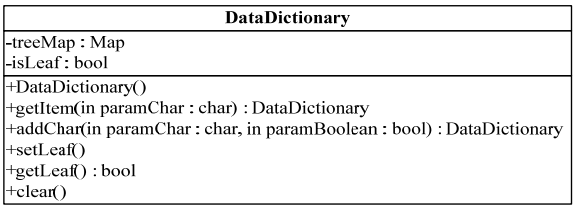


Figure 7 The data dictionary class diagram

2) The communication module

This is one of system cores, responsible for the reliable transfer of data among various nodes and triggers the data processing in the business layer. The module contains two abstract classes and a helper class.

3) Chinese word segmentation module

Chinese word segmentation is a key task in the system that makes metadata extraction possible. Current system adopts the longest string matching approach. The module consists of two classes: MyBaseAnalyzer and SentencesAnalyzer.

4) The CN module

The main class in the module is the CenterMian class whose main() method will load system parameters via

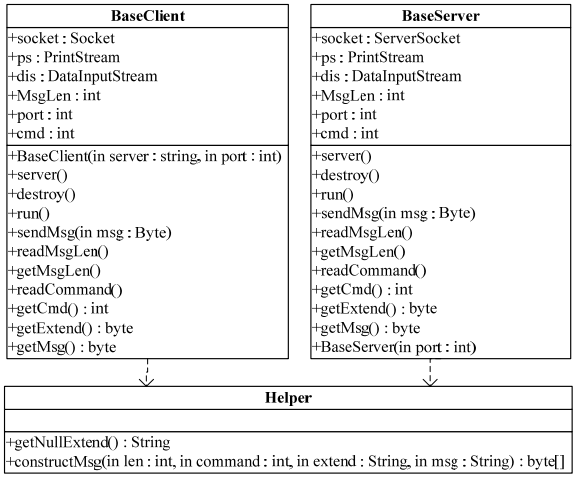


Figure 8 The communication module class diagram

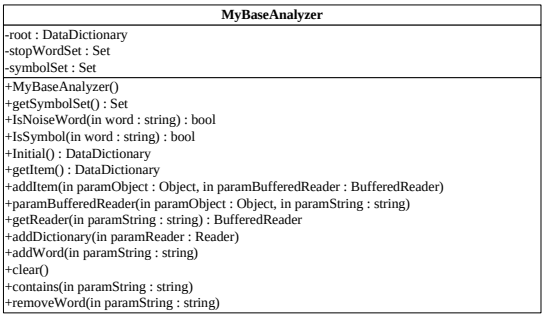


Figure 9 The MyBaseAnalyzer class diagram

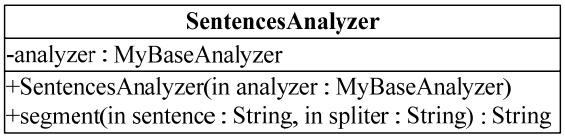


Figure 10 The SentencesAnalyzer class diagram

class ConfigUtil and will create an instance of CenterServer, then start it.

5) The Client module

This module consists of functionality class such as BrowserClient, PdfboxParse, CenterRouter, Command, DestInfor, NormalMeta and SearchResult.

D. Experiments

A prototype Paper Sharing System has been developed and deployed for system test. The aim is to verify

1) The construction of SPN network. Configure the number of SPN to be 4 and register them to check if the routing table in CN is correct.

2) The maintenance of SN network. Start SPN, join some SN to each SPN, then, detach some SN, and join the exited SN to SPN again, to check if the routing table in the SPN is correct.

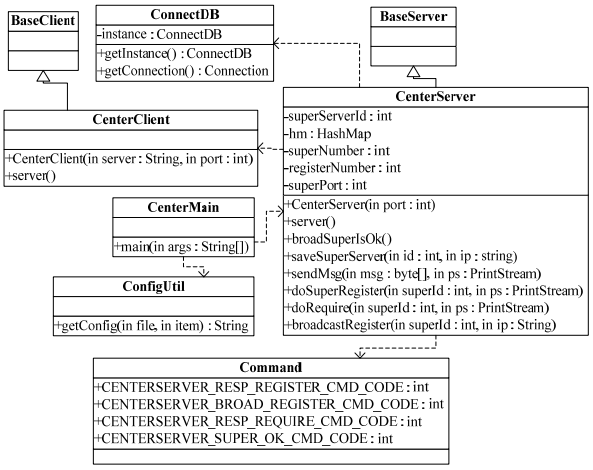


Figure 11 The CN module class diagram

3) Search of papers. Search papers via a browser to check if papers can be located by title or metadata information correctly.

4) Prevention of single point of failure. When an SN failed, its resources are transmitted to its successor and are still available.

The system is deployed in a local network with 2 SPN and 4 SN which is depicted in Fig.13. The test cases listed above are run and results are recorded as below.

1) The construction of SPN network.

Start the service in CN with ip 125.216.250.70, then, start the services in two SPN with ip 125.216.249.109 and 125.216.249.200 respectively. The CN routing table reads as

TABLE 1 THE CN ROUTING TABLE

Ip	superId
125.216.249.200	1
125.216.249.109	2

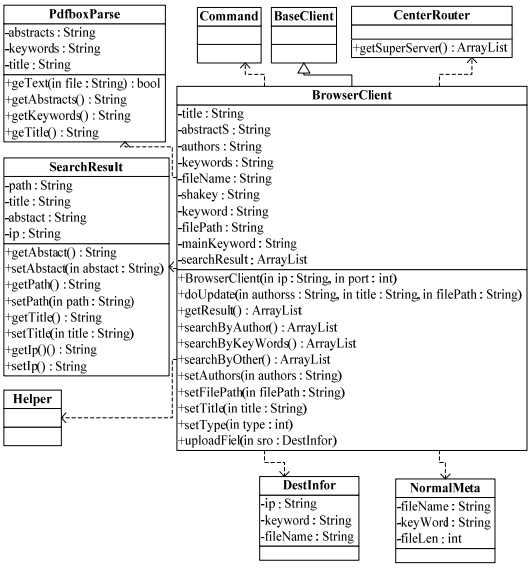


Figure 12 The Client module class diagram

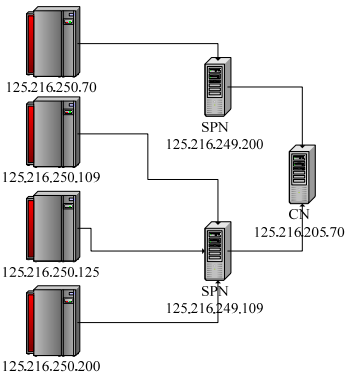


Figure 13 The topology of the P2P server system

2) The maintenance of SN network.

Add an SN to SPN with superId 1, and 3 SN to SPN

TABLE 2 THE SPN ROUTING TABLE

Ip	status	superId	hashkey
125.216.250.70	0	1	579269F26FA29E8B57E35 3D791A582E7E749A529
125.216.250.109	0	2	48F2DB51CC5ED84A84E7 E1A7F4EFF8BE17CBD543
125.216.250.125	0	2	117FA19443F995E6C73A2 91933974CB0C9584282
125.216.250.200	0	2	6EA5FAFF06F00F5256942 14B45EEB7D33274BB11

with superId 2, the SPN routing table reads as

Shutdown the SN with ip 125.216.250.70, the corresponding status in Table 2 turns to be 1; restart it again, the status changes back to 0.

3) Search of papers. Search by title is carried out as blow.

Figure 14 Search by title

4) Prevention of single point of failure. First search by key word and verify that a paper containing the key word does exist in some SN, see Fig. 15, then, shutdown the SN and search again, the paper still exists in the network, although in another SN (refer to Fig. 16).

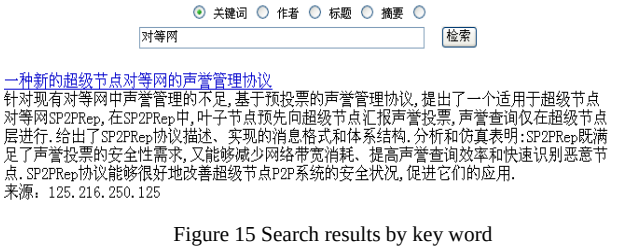


Figure 15 Search results by key word

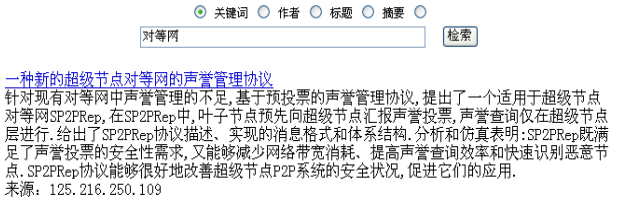


Figure 16 Search results by key word when the SN containing the resource is shutdown

From the above experiments it can be concluded that all the test tasks are fulfilled as desired.

IV. CONCLUSIONS

P2P network is a promising technology for resource sharing system for group/team/organizations. Most of current systems suffer high bandwidth pressure and/or low search efficiency. In this paper, a layered P2P server system model based on the Chord and SHA algorithm is proposed to resolve the high bandwidth problem. The introduction of CN eliminates unnecessary duplicate resources storage in the network; the introduction of SPN eliminates the flooding of routing information around the network. By incorporating Chinese word segmentation and metadata extraction modules, the system supports versatile and convenient search approaches, thus enhancing search efficiency. Furthermore, the design and implementation of a P2P Paper Sharing System based on the new model is discussed. Experimental results demonstrate that the system achieves its design goals.

There remain some topics uncovered by the present paper. Semantic search, for example, is an important issue that is not discussed here and deserves further study. These will be addressed in subsequent papers.

ACKNOWLEDGMENT

The authors would like to thank the Guangdong Provincial Lab for Fundamental Software and Application Construction Technology for its support in preparing the experimental environment.

REFERENCES

[1] Dejan S. Milojicic, et al. Peer-to-Peer Computing, Hewlett-Packard Company, 2002.

[2] I. Stoica, R. Morris R, and D. Karger, Chord: a scalable Peer-to-Peer Lookup service for Internet Applications, Proceedings of ACM SIGCOMM 2001.San Diego CA, pp. 149-160.

[3] Huachun LIU, Sort and Key Technique Analysis of the P2P Networks, Microcomputer Information, 2008, No. 9, pp. 112-114.

[4] Napster [EB/OL]. http://www.napster.com.

[5] U. Lechner, and B. Schmid, Communities - Business models and system architectures: The blueprint of MP3.com, Napster and Gnutella revisited, Proceedings of the Hawaii International Conference on System Sciences, p 164, 2001.

[6] KaZaA [EB/OL]. http://www.kazaa.com.

[7] Rowstron A, Druschel P. Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems[J]. IFIP/ACM International Conference on Distributed Systems Platforms. Heidelberg. 2001: 329-350.

[8] Hua Chen, Mao Yang. Maze:A Social Peer-to-Peer Network[J]. Proc.of the International Conference on E-commerce Technology for Dynamic E-business. Beijing, China, IEEE Press, 2004.

[9] Granary[EB/OL]: http://hpc.cs.tsinghua.edu.cn/granary/.

[10] B. Yang, and H. Garcia-Molina, Improving Search in Peer-to-Peer Networks, ICDCS 2002, pp. 5-14.

[11] Libin LIANG, Jun ZHANG, and Xilin LUO, P2P algorithm based on network proximity, Computer Engineering and Design, 2005, No. 9, pp. 2308-2311.

[12] Q. Lv, P. Cao, F. Cohen, et al. Search and Replication in Unstructured Peer-to-Peer Networks, ICS, 2002.

[13] Y. Chawathe, and S. Rantnasamy, Making Gnutella-like P2P System Scalable, SIGCOMM, 2003.

[14] Yunhao LIU et al. Location-Aware Topology Matching in P2P Systems, Proceedings of IEEE INFOCOM, 2004.

[15] D. Tsoumakos et al. A Comparison of Peer-to-Peer Search Methods, WebDB, 2002.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 365-368

Determinant Quantum Key Distribution Via Entanglement Swapping

Nanrun Zhou, Xiawen Xiao, Lijun Wang, and Lihua Gong
Department of Electronic Information Engineering, Nanchang University, Nanchang, China
Email: znr21@163.com (NR Zhou)

***Abstract*—The efficiency of the quantum key agreement protocol that was designed by replacing classical channel with a quantum one during quantum teleportation is reanalyzed and an improved quantum key distribution protocol is presented by making best of the properties of entanglement swapping. The improved quantum key distribution protocol is secure because the key bits only depend on the random measurement results and are independent of the states when transmitting over the channel, moreover, the eavesdropping can be detected by comparing a subset of the resulting key bits.**

***Index Terms*—quantum key distribution, teleportation, entanglement swapping, information security**

I. Introduction

Quantum key distribution (QKD) or quantum key agreement (QKA) is a technique exploiting the fundamental laws of quantum mechanics to obtain secret key with provable security. The goal of QKD protocols is to create private key between two parties or even more over a public or unsecured channel. The only requirement for QKD protocols is that quantum information can be communicated over the public channel with an error rate lower than a certain threshold [1]. The resulting key can then be used to implement a classical private key cryptosystem, which enables the parties to communicate securely [2,3].

Since the initial proposal of QKD in 1984 (BB84 protocol) by Bennett and Brassard [4], a number of researchers have devoted to this field [5~7]. Experimentally, much progress on QKD has been made both in free-space systems and in optical fiber systems [8~27]. In 2000, the QKD system over a 48km optical fiber network was reported by R. G. Hughes, G. L. Morgan and C. G. Peterson [12]. In 2002, the 67km optic-fiber QKD experiment was performed successfully by Geneva University [13]. QKD over 14.8 km in a special optical fiber was accomplished by Guo's group in 2003 [14]. In the year of 2004, the distance of QKD was extended greatly. QKD in 50km optic fibers was reported to be stable over six hours and a practical quantum communication in a local area network using the TCP protocol was demonstrated by Zeng's group [15]. Single-photon interference experiment over 100 km for quantum cryptography system using balanced gated-mode photon detector was accomplished by Japanese researchers [16]. S. Fasel and N. Gisin et al demonstrated solutions to the chromatic dispersion issue for an energy-time entanglement based on QKD system using over 30km standard fiber quantum channel [17]. A short wavelength gigahertz clocked fiber-optic QKD system was developed using a standard telecommunications optical fiber, which can be applied to short distance QKD, such as campus- or metropolitan-scale networks [18]. Quantum key distribution over 122km of standard telecom fiber was achieved by C. Gobby, Z. L. Yuan, and A. J. Shields [19]. The entangled photons were distributed directly through the atmosphere to a receiver's station 7.8 km away at night [20]. A reverse-reconciliated coherent-state continuous-variable quantum key distribution system was implemented over 25 km optical fiber in 2007 [21]. The free-space decoy-state or entanglement-based quantum key distribution over 144 km was demonstrated experimentally [22,23]. An entangled QKD system over two free-space optical links producing a total separation of 1,575 m was realized in 2008 [24]. T. Honjo et al reported the first entanglement-based QKD experiment over a 100 km optical fiber in the same year [25]. Feasibility of QKD over distances of 300 km or even more with entangled state sources placed in the middle between Alice and Bob was confirmed in 2009 [26]. These experimental results indicate that quantum cryptography and quantum communication can be put into real-life use technically. Actually, commercial quantum cryptography device has been available already [27].

The goal of this paper is to analyze the efficiency of the QKA protocol [7] we proposed in the year of 2004, and thus to present an improved QKD protocol. This paper is organized as follows. We analyze the efficiency of the QKA protocol based on quantum teleportation in Section 2. In Section 3, we propose an improved QKD protocol based on quantum entanglement swapping (QES). Finally, we make a brief conclusion in Section 4.

II. Efficiency Analysis of QKA Protocol Based on Quantum Teleportation

In this section, we will analyze the efficiency of the QKA protocol [7] based on quantum teleportation from several aspects. The QKA protocol is simpler than the BB84 protocol [4], EPR protocol [5] and B92 protocol [6] in a way. In the following, we take BB84 protocol as a comparison.

Supported by the National Natural Science Foundation of China (10647133), the Natural Science Foundation of Jiangxi Province (2007GQS1906), the Research Foundation of the Education Dept. of Jiangxi Province ([2007]22), and the Key Project in the 11th Five-Year Plan on Education Science of Jiangxi Province (07ZD017)

Firstly, BB84 protocol requires Alice to choose two strings of true random bits and to encode one string of random bits in different bases according to the other string of random bits. It also requires Bob to measure each qubit in two different bases at random. Thus, BB84 protocol needs three strings of true random bits. However, the QKA protocol needs no random bits because the random key bits come from the randomness of the measurement results. As is known, the generation of true random bits is not an easy job. From this point of view, the QKA protocol needs lower cost.

Secondly, in the case of no eavesdropping, the rate of key bits to quantum states of the QKA protocol is as high as that of BB84 protocol or EPR protocol. To obtain n key bits, both BB84 protocol and EPR protocol need at least $2n$ quantum states. To get $2n$ key bits, the QKA protocol consumes at least $4n$ quantum states. Therefore, the rate of key bits to quantum states of the QKA protocol, BB84 protocol and EPR protocol is 50% in the absence of Eve.

Finally, in BB84 protocol, Alice and Bob communicate classically their choice of bases and discard all measurement results where different bases were used. Furthermore, to detect the existence of Eve, BB84 protocol requires Alice to select a subset of the key bits they generated; that is to say, it must compromise a subset of the key bits even there is no interference from Eve. While the QKA protocol can keep all the key bits if there is no interference from Eve because the existence of Eve can be detected according to the validity of the results Alice obtained. In other words, the QKA protocol need not compromise a subset of the key bits to check on Eve's interference. So the rate of key bits to quantum states of the QKA protocol is necessarily higher than that of BB84 protocol.

In brief, the QKA protocol designed by replacing classical channel with a quantum one during quantum teleportation has its advantages in efficiency. Of course, it also has some disadvantages which will be discussed next section.

III. DETERMINANT QUANTUM KEY DISTRIBUTION PROTOCOL BASED ON ENTANGLEMENT SWAPPING

Above QKA protocol is actually based on quantum teleportation, which requires Bob to send a series of photons in the states of same to the corresponding states teleported and their amplitudes to Alice. Alice must perform certain operation in order to get the key bits. Briefly, the QKA protocol requires more classical communications than BB84 protocol. Motivated by this, we modify the QKA protocol by employing quantum entanglement swapping. Thus local operation and classical communication of the improved QKD protocol are considerably reduced.

A. The improved QKD protocol

The QKD protocol based on entanglement swapping (ES) is illustrated in Fig.1. Alice possesses entangled state containing photons 1 and 2, and photon 3 entangled with Bob's photon 4.

Alice performs an ES operation on photon 2 and photon 3, photon 1 and photon 4 will be entangled. And the whole system containing photons 1, 2, 3 and 4 can be rewritten as:

$$\begin{aligned}|\psi\rangle_{1234} &= |\psi^-\rangle_{12}\otimes|\psi^-\rangle_{34}\\ &= \frac{1}{\sqrt{2}}\left(|0\rangle_1|1\rangle_2-|1\rangle_1|0\rangle_2\right)\otimes\frac{1}{\sqrt{2}}\left(|0\rangle_3|1\rangle_4-|1\rangle_3|0\rangle_4\right)\\ &= \frac{1}{2}\left[|\psi^+\rangle_{23}|\psi^+\rangle_{14}-|\psi^-\rangle_{23}|\psi^-\rangle_{14}-|\phi^+\rangle_{23}|\phi^+\rangle_{14}+|\phi^-\rangle_{23}|\phi^-\rangle_{14}\right]\end{aligned} \quad (1)$$

From (1), it is easy to see that Alice's photon 1 and Bob's photon 4 will end up in one of the following four Bell states $|\psi^+\rangle_{14}$, $|\psi^-\rangle_{14}$, $|\phi^+\rangle_{14}$ and $|\phi^-\rangle_{14}$ with equal probability $1/4$, corresponding to Alice's measurement outcomes $|\psi^+\rangle_{23}$, $|\psi^-\rangle_{23}$, $|\phi^+\rangle_{23}$ and $|\phi^-\rangle_{23}$, respectively. For example, if Alice's measurement result is $|\psi^+\rangle_{23}$, then photon 1 and photon 4 must be in the state $|\psi^+\rangle_{14}$. If photons 1 and 4 are in the states $|\psi^-\rangle_{14}$ or $|\psi^+\rangle_{14}$, then the measurement results of photon 1 and photon 4 will be different; if photon 1 and photon 4 are in the states $|\phi^-\rangle_{14}$ or $|\phi^+\rangle_{14}$, then the results of photon 1 and photon 4 will be identical. Bob measures photon 4 and records the result as a key bit. Alice records the result of photon 1 as a key bit if photons 2 and 3 are in the state $|\phi^-\rangle_{23}$ or $|\phi^+\rangle_{23}$. Otherwise, Alice takes the NOT of the result of photon 1 as a key bit if her result of photons 2 and 3 is $|\psi^+\rangle_{23}$ or $|\psi^-\rangle_{23}$ (see Table 1).

Based on the above properties of ES, we present an improved QKD protocol. The improved QKD protocol works as follows.

(1) Alice prepares two entangled pairs, one of which contains photons 1 and 2 in the state $|\psi^-\rangle_{12}$ and the other contains photons 3 and 4 in the state $|\psi^-\rangle_{34}$, and then sends photon 4 to Bob.

(2) Alice measures photons 2 and 3 in the Bell states, then photons 1 and 4 must be entangled. Alice measures photon 1 in the X basis, with the knowledge of photons 2 and 3, Alice can infer the result of Bob's photon 4, and records it as a key bit.

(3) Bob measures photon 4 in the X basis and records the result as a key bit.

(4) Repeating above steps, Alice and Bob would get two strings of key bits.

(5) Alice and Bob choose a subset of the obtained key bits to detect whether there is an eavesdropper. If the

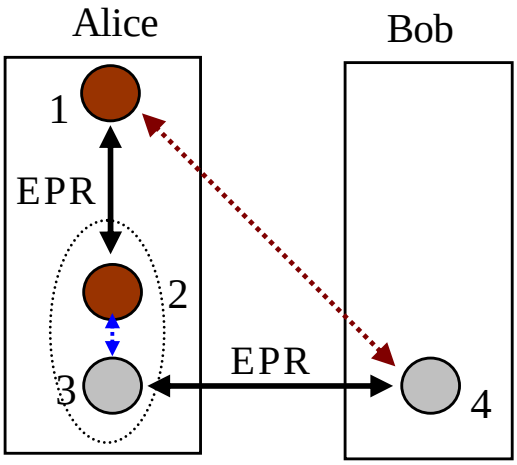


Figure 1. Schematic of quantum key distribution based on entanglement swapping.

error rate is acceptable, they keep the two remaining strings to generate a same key. Otherwise, abolish the results and turn to a new round of QKD.

(6) Alice and Bob perform information reconciliation and privacy amplification on the remaining key bits to obtain finally shared secure key bits.

When there is no eavesdropping, through sending three photons and two amplitudes α and β of the photon to be teleported and performing two local measurements, the QKA protocol based on quantum teleportation can generate two key bits averagely. While by sending just one photon and performing three local measurements, the improved QKD protocol via ES can create one key bit in average. Therefore, the improved QKD protocol needs less quantum resources and less classical and quantum communications.

The BB84 protocol [4], EPR protocol [5] and B92 protocol [6] can all generate a secure random key, however, they are non-deterministic. While in the improved QKD protocol, Alice can infer Bob's measurement results with the measurement outcomes of photons 2 and 3 before the eavesdropper is detected, and then Alice and Bob do not need to compare their own measurement bases or discard the resulting bits, which are the advantages of the improved QKD protocol. From the QKD process, it is easily known that the final key bit shared by Alice and Bob is determined by the measurement result of photons 2 and 3 to a certain extent. That is to say, the improved QKD protocol can distribute a key in a determinant way, which is the main advantage of our improved QKD protocol.

B. Security Analyses

The QKD protocol is quite secure for several reasons. Firstly, the photons Alice sends to Bob are basically useless before Alice performs ES operation. Secondly, we analyze the intercept/resend strategy. When Alice sends the photon 4 to Bob, an eavesdropper Eve may intercept this photon. Before Alice performs an ES operation on photon 2 and photon 3, the state of photon 4 intended for Bob is

TABLE 1
RELATIONSHIP BETWEEN OUTCOMES AND KEY BITS

Bob's result	Photon 1	Photons 2 and 3	Alice's bit
0	0	$\left\|\phi^{\pm}\right\rangle_{23}$	0
	1	$\left\|\psi^{\pm}\right\rangle_{23}$	
1	0	$\left\|\psi^{\pm}\right\rangle_{23}$	1
	1	$\left\|\phi^{\pm}\right\rangle_{23}$	

$$\rho_4 = \mathrm{Tr}_3 \rho_{\left|\psi^-\right\rangle_{34}} = \frac{1}{2}\left(\left|0\right\rangle\left\langle 0\right| + \left|1\right\rangle\left\langle 1\right|\right). \quad (2)$$

From (2), it is easy to see that photon 4 is in a mixed state. After Alice performs an ES operation on photon 2 and photon 3, photon 1 and photon 4 are entangled in one of the following four Bell states $\left|\psi^+\right\rangle_{14}$, $\left|\psi^-\right\rangle_{14}$, $\left|\phi^+\right\rangle_{14}$ and $\left|\phi^-\right\rangle_{14}$ with equal probability $1/4$. At this time, the state of photon 4 is still mixed, i.e., $\rho_4 = \frac{1}{2}\left(\left|0\right\rangle\left\langle 0\right| + \left|1\right\rangle\left\langle 1\right|\right)$. Thus Eve can not obtain any information by this way even though the intercepted photon has been measured by Eve. If Eve resends a fake photon in the state $\left|\varphi\right\rangle_E = \alpha\left|0\right\rangle + \beta\left|1\right\rangle$ to Bob, where $\left|\alpha\right|^2 + \left|\beta\right|^2 = 1$. After receiving this photon, Bob measures this fake photon in the X basis. Then, Bob obtains the measurement outcome 0 with probability $\left|\alpha\right|^2$ and the measurement outcome 1 with probability $\left|\beta\right|^2$, respectively. Thus, the error rate must rise and is easy to exceed the threshold, and then Alice and Bob abort this round communications. Besides, Eve somehow eavesdropped on the classical communication channels through which Alice and Bob choose a subset of key bits to check on Eve's interference, she would still be caught if the identity of Bob is verified. Finally, it is conceivable that an eavesdropper might obtain partial information by entangling enough ancillary photons with the photons sent to Bob, but if Alice distributes a series of photons to Bob and such entanglement could be detected by tests conducted on "sample" EPR entangled states before secure QKD. If the "sample" EPR entangled states are altered or destroyed, they abort this round communications.

IV. CONCLUSION

In this paper, we analyzed the efficiency of the QKA protocol based on quantum teleportation and presented an

improved QKD protocol based on quantum entanglement swapping. It is shown that the efficiency of the QKA protocol is higher than that of BB84 protocol. And the improved QKD protocol can accomplish secure QKD with less classical communications and quantum resources. Both protocols have their own advantages, so we can choose them for different purposes. For example, the improved QKD protocol just requires Bob to measure the photons sent from Alice in the X basis, so it is convenient for the case that Bob has no much ability to manipulate quantum states. Unlike the non-deterministic protocols such as BB84 protocol [4], EPR protocol [5] and B92 protocol [6], the final key shared by Alice and Bob can be distributed in a determinant way, because the final key is determined by the measurement result of photons 2 and 3.

QKD is of particular importance in the field of quantum cryptography, among which multiparty QKD is a new focus for its great use in group communications, such as multiparty computations, network games and video conferences [28]. The improved QKD protocol can be used to construct multiparty QKD and the relevant results will be published elsewhere.

REFERENCES

[1] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information.* Cambridge University, 2000.

[2] N. R. Zhou and G. H. Zeng, "A realizable quantum encryption algorithm for qubits," *Chin. Phys.*, vol.14, pp. 2164–2169, 2005.

[3] N. R. Zhou, G. H. Zeng, Y. Y. Nie, J. Xiong and F. C. Zhu, "A novel quantum block encryption algorithm based on quantum computation," *Physica A,* vol. 362, pp. 305–313, 2006.

[4] C. H. Bennett and G. Brassard, "Quantum cryptography: public key distribution and coin tossing," in *Proceedings of the IEEE International Conference on Computers, Systems and Signal Processing, Bangalore (IEEE, New York)* 1984, pp. 175–179.

[5] A. K. Ekert, "Quantum cryptography based on Bell's theorem," *Phys. Rev. Lett.*, vol. 67, pp. 661–664, 1991.

[6] C. H. Bennett, "Quantum cryptography using any two nonorthogonal states," *Phys.* Rev. *Lett.*, vol. 68, pp. 3121–3124, 1992.

[7] N. R. Zhou, G. H. Zeng and J. Xiong, "Quantum key agreement protocol," *Electron. Lett*., vol. 40, pp. 1149–1150, 2004.

[8] R. J. Hughes, J. E. Nordholt, D. Derkacs and C. G. Peterson, "Practical free-space quantum key distribution over 10 km in daylight and at night," *New. J. Phys.*, vol. 4, pp. 43.1–43.14, 2002.

[9] C. Kurtsiefer, P. Zarda, M. Halder, H. Weinfurter, P. M. Gorman, et al., "Quantum cryptography: A step towards global key distribution," *Nature*, vol. 419, pp. 450, 2002.

[10] C. Marand and P. D. Townsend, "Quantum key distribution over distances as long as 30 km," *Opt. Lett.*, vol. 20, pp. 1695–1697, 1995.

[11] P. A. Hiskett, G. Bonfrate, G. S. Buller and P. D. Townsend, "Eighty kilometre transmission experiment using an InGaAs/InP SPAD-based quantum cryptography receiver operating at 1.55 km," *J. Mod. Opt.*, vol. 48, pp. 1957–1966, 2001.

[12] R. J. Hughes, G. L. Morgan and C. G. Peterson, "Quantum key distribution over a 48km optical fibre network," *J. Mod. Opt.*, vol. 47, pp. 533–547, 2000.

[13] D. Stucki, N. Gisin, O. Guinnard, G. Ribordy and H. Zbinden, "Quantum key distribution over 67km with a plug&play system," *New J. Phys.*, vol. 4, pp. 41.1–41.8, 2002.

[14] Y. Z. Gui, Z. F. Han, X. F. Mo and G. C. Guo, "Experimental quantum key distribution over 14.8 km in a special optical fibre," *Chin. Phys. Lett.*, vol. 20, pp. 608–610, 2003.

[15] C. Y. Zhou, G. Wu, X. L. Chen, H. X. Li, and H. P. Zeng, "Quantum key distribution in 50-km optic fibers," *Science in China Series G: Physics, Mechanics and Astronomy*, vol. 47, pp. 182–188, 2004.

[16] H. Kosaka, A. Tomita, Y. Nambu, T. Kimura and K. Nakamura, "Single-photon interference experiment over 100 km for quantum cryptography system using balanced gated-mode photon detector," *Electron. Lett.*, vol. 39, pp. 1199–1201, 2003.

[17] S. Fasel, N. Gisin, G. Ribordy and H. Zbinden, "Quantum key distribution over 30 km of standard fiber using energy-time entangled photon pairs: a comparison of two chromatic dispersion reduction methods," *Eur. Phys. J. D*, vol. 30, pp. 143–148, 2004.

[18] K. J. Gordon, V. Fernandez, P. D. Townsend and G. S. Buller, "A short wavelength GigaHertz clocked fiber-optic quantum key distribution system," *IEEE J. Quantum Electron.*, vol. 40, pp. 900–908, 2004.

[19] C. Gobby, Z. L. Yuan and A. J. Shields, "Quantum key distribution over 122 km of standard telecom fiber," *Appl. Phys. Lett.*, vol. 84, pp. 3762, 2004.

[20] K. J. Resch, M. Lindenthal, B. Blauensteiner, H.R. Böhm, A. Fedrizzi, et al., "Distributing entanglement and single photons through an intra-city, free-space quantum channel," *Opt. Expr.*, vol. 20, pp. 202–209, 2005.

[21] J. Lodewyck, M. Bloch, R. García-Patrón, S. Fossier, E. Karpov, et al., "Quantum key distribution over 25 km with an all-fiber continuous-variable system," *Phys.* Rev. *A*, vol. 76**,** pp. 042305.1–042305.10, 2007.

[22] T. Schmitt-Manderbach, H. Weier, M. Furst, R. Ursin, F. Tiefenbacher, et al., "Experimental demonstration of free-space decoy-state quantum key distribution over 144 km," *Phys. Rev. Lett.*, vol. 98, pp. 010504.1–010504.4, 2007.

[23] R. Ursin, F. Tiefenbacher, T. Schmitt-Manderbach, H. Weier, T. Scheidl, et al., "Entanglement-based quantum communication over 144km," *Nature Physics*, vol. 3, pp. 481–486, 2007.

[24] C. Erven, C. Couteau, R. Laflamme and G. Weihs, "Entangled quantum key distribution over two free-space optical links," *Opt. Expr.*, vol. 16, pp. 16840–16853, 2008.

[25] T. Honjo, S. W. Nam, H. Takesue, Q. Zhang, H. Kamada, et al., "Long-distance entanglement-based quantum key distribution over optical fiber," *Opt. Expr*. vol. 16**,** pp. 19118–19126, 2008.

[26] T. Scheidl, R. Ursin, A. Fedrizzi, S. Ramelow, X.S. Ma, et al, "Feasibility of 300km quantum key distribution with entangled states," *New. J. Phys.*, vol. 11, pp. 085002.1–085002.13, 2009.

[27] id Quantique S A http://www.idquantique.com

[28] N. R. Zhou, Design and analyses of quantum secure communication protocols. Doctoral dissertation, Shanghai Jiaotong University, 2005.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 369-372

Research on Layout Algorithms for Better Data Visualization

Luhe Hong, Fanlin Meng, and Jianli Cai
Department of Automation, Xiamen University, Xiamen, China
Email: hongluhe@163.com, manard@126.com, cjl701@sina.com

***Abstract*—Based on the theory and technology of data visualization, this paper improves some layout algorithms regarding its own natures and characteristics. It also discusses the problem of how to choose proper layout algorithm for specific data to make the visualization better. Using the improved algorithms, we can simply convert a series of complex data into more expressive images as to convey the important information hidden within the data more effectively and accurately.**

***Index Terms*—data visualization, layout algorithm, chart, information**

I. Introduction

Data visualization derived from visualization in scientific computing. People express a variety of statistical data via tables, curves, charts and other means at the earliest [1]. Data visualization refers to the use of computer graphics and image processing technologies to convert the data into displayed graphic or image for further interactive processing, which helps to convey the information more clearly and effectively. It covers a number of areas including computer graphics, image processing, computer-aided design, computer vision, human-computer interaction technology and other fields. With the rapid development of computer technology, application areas of visualization have also been widened. Many research institutes and companies at home and abroad have made a lot of achievements in this area.

Nowadays, the science and technology have developed rapidly. A large number of diverse and complex information produced in the real world, most of which is in the form of tables stored in the database both dull and difficult to understand. Then how to enable users discover the important hidden information within the data is a serious problem. Data visualization technology provides an effective analysis channel to solve this problem. It can analyze the original data files to gain the useful information, and express the results in graphic or image that shows the association, comparison and trends among the data, which does help users understand a large number of complex and abstract data information in an efficient and accurate way.

II. Directed Graph of Layout Algorithm

A. The Definition of A Directed Graph

Drawing directed graphs for layout algorithms is an important data visualization technology[2][3]. Take every entity as a node in the directed graph and the dependency relationship between the entities can be expressed by the edge between nodes. Edge direction is determined by the dependency relationship, i.e., if entity A use the service provided by entity B, then we say entity A depends on entity B and there exists a directed edge from the node on behalf of entity A to the node on behalf of entity B.

B. The Main Criteria for Layout Algorithm

Aesthetic is one of the important criteria that measure the strength and weakness of layout algorithm [4]. In order to make the layout results more readable, it needs to meet some certain aesthetic criteria, such as a clear hierarchical structure, the graphics as even and symmetry as possible. However, in our design of layout algorithms we primarily consider these following factors:

- It concerns about the space utilization ratio of the layout under the same zoom. The higher the utilization ratio is, the smaller the drawing area occupies. Thus, the layout will show more compact, and is easier for users to observe the layout in the whole. In addition, it reduces the frequency of switching layout view. It is especially important in large scale system with a great amount of data.
- The number of crossing edges is also taken into consideration. The smaller the number of crossing edges is, the clearer the relations between the data are.
- It also concerns about the clarity on displaying the aggregation among entities, namely, it matters whether a group of entities depend on each other are well displayed in e layout process. If so, it will be convenient to analysis these interdependent data.

Generally speaking, it is very difficult to satisfy all of the above criteria at the same time, as some criteria themselves are NP-complete problems, such as minimizing the number of crossing edges. Therefore, the general layout algorithms will place extra emphasis on one or several criteria based on their own concerns.

III. Improvements on Common Layout Algorithms and Special Applications

The following aims at five different kinds of common layout algorithms. We will make specific analysis and comparison respectively, improve each algorithm

Corresponding author: Jianli Cai
E-mail: cjl701@sina.com

AP-PROC-CS-09CN005

appropriately with its own respective nature, and then discuss the applied fields.

A. The Circular Layout Algorithm

Circular layout is the most prominent and oldest conventions used to draw graphs. Circular graph layout is a drawing scheme where all nodes are placed on the perimeter of a circle. In such a layout, the edges connecting these nodes are passing within the circle. In particular, a circular layout is appropriate for applications that emphasize the clustering decomposition of a graph, where each cluster is drawn on a separate circle. Much work has been done on these layouts, most of it addressing both the layout of a single circle as well as positioning multiple circles together, in order to illustrate clearly the various clusters composing the full graph[5][6].

The improved circular layouts can find a part of the entities that might be isolated into small circular by re-arranging the source data and a series of intelligent detections. By the same means, we can find a few focal points that have more frequently links from the remaining entities and evenly distribute them on the circle. Thus, we need to handle the positions of entities connected with the focal points to make these focal points be clearly displayed in the graph while the cross between connections can be reduced. The circular layout is mainly used for small-medium-sized data analysis, which combines the main advantages of hierarchical layout and radial layout. As a result, we got high-level layout space utilization in the same proportion, with clearer hierarchy structure and less cross-borders, and the clarity of clustering performance is better.

This drawing convention is often used for the layout of networks and systems management diagrams, where it naturally captures the essence of ring and star topologies. It can be also used for other kinds of graphs, such as social networks and WWW graphs. An inherent issue with circular layouts is that the rigid restriction on node placement often gives rise to long edges and an overall dense drawing. The graph of circular layout for 2000 e-mail records of is shown as Figure 1.

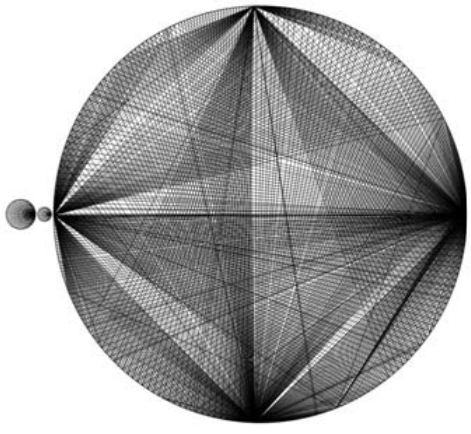


Figure 1. The graph of circular layout for 2000 records of e-mail

B. The Tree Layout Algorithm

The classic algorithm proposed and improved by Reingold and Tilford[7][8], gives a very effective solution to produce a classical, top–down drawing of rooted trees. The classical, top–down drawing of trees has huge advantages and was widely used in many applications. Its use for displaying hierarchies benefits from its natural interpretation.

In the improved tree layout algorithm, the user can designated an entity as root node according to known information. Then, the layout will be carried out, while each sub-node of the root node will be the parent node of the next level nodes. The clarity of hierarchy structure is clear so that it becomes easy to view the association properties of special entities.

This algorithm is broadly divided into four steps. Firstly, we should determine the layers which each node belongs to according to the direction of edges. Then, adjust the order of nodes in each layer in order to reduce cross-border. After that, we need to adjust the location of nodes in each layer to shorten the length of the edge and finally we should draw the edges. Computational complexity of the algorithm is O (n^2). The advantages of this algorithm are shown as following: clarity of hierarchy structure of dependency relationships and less cross-border. However, the drawback is that when the number of nodes is too large, the drawing area occupied will be relatively large, which is inconvenient to display all nodes in overall view. The tree layout is suitable for large and medium-sized data, and the data need to have a clear hierarchy. The graph of tree layout for 1000 chat log records of MSN is shown as Figure 2.

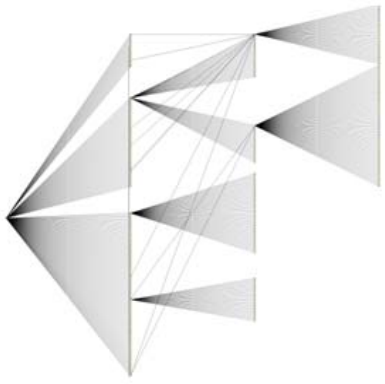


Figure 2. The graph of tree layout for 1000 chat log records of MSN

C. The Radial Layout Algorithm

The main idea of radial layout algorithm described as follow[9]: Given a focal point for A, and any node R, the structure of spanning tree needs to meet that the conditions that the distance from A to R in the tree should be the shortest path among each two points in the graph. This algorithm is fast, but there is a drawback that the effectiveness of the layout algorithm over-dependent on the selection of center point. If the user does not give the center point, then the layout algorithm will select the largest node that has the shortest distance to the leaf node

as the center point. If there is no leaf node, then randomly select a node as a center point. Therefore, radial layout algorithm is particularly suitable for the tree structure with large number of nodes and the smaller density of edges. It has high-level layout space utilization, clearer display of aggregation relationship, but more cross-borders.

The improved radial layout will re-sort out the complex group of interconnected entities in the data to highlight their mesh structure. The group settings of an entity are linked with the special entity. The edges linking to this single entity seem to be a beam of light radiating out from the position of the entity. The advantages of this layout are: the level of hierarchy structure is clear, the group operation of the data is evident, and little possibility to generate excessive crosses between entities. When there are many links between entities in the data, the radial layout becomes quite useful. The radial layout is suitable for large and medium-sized data analysis, for example: identify the center with complex and large number of acts in a large database, typically telephone message declarations, e-mail between records and financial transactions, etc. The graph of radial layout for 500 records of telephone call is shown as Figure 3.

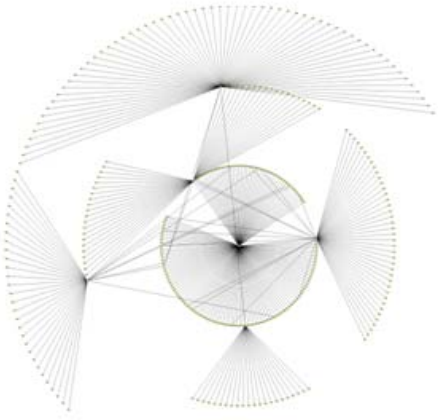


Figure 3. The graph of radial layout for 500 records of telephone call

D. The Balloon Layout Algorithm

Balloon layout is a special kind of layout, which is based on the improved tree layout [10]. The balloon layout as the tree layout, it also supports the high fault tolerance of the data, and supports multi-tree display. A balloon drawing of a rooted tree is a drawing having the following properties:

- All the children under the same parent are placed on the circumference of the circle centered at their parent.
- There exist no edge crossings in the drawing.
- For the two edges on any path from the root node, the farther from the root an edge is, the shorter its drawing length becomes.

Cluster is the main technology of balloon layout and is an important means of knowledge discovery in data mining. It can extract the information from the vast amounts of data that implicit, previously unknown, and there is potential value in decision-making. When the mining task face the data set of lacking area knowledge or the knowledge is not complete, cluster analysis technique can automatically divide the unmarked data objects into different classes, and without the constraint and interference of the prior knowledge of people, and gain the information exists in the data set originally.

The improved layout of the balloon arranges all the child nodes connected to the parent node at each reasonable position of concentric circle one by one, and then forms a few aggregation graphs, so that can show the different groups of the interconnection entities and the relationship between them. Aggregation allows entities into different groups, so that can identify the entities in two or more groups, the group of interconnect entities is displayed in the center of the chart, while other groups disperse from the center group. This layout mean will minimize the data cross between each concentric circle, it is easy for people to find the center, so that is useful for the data contains a lot of links, therefore particularly suitable for officers to analysis the communication information and financial transactions of the suspect and so on. The graph of balloon layout for 1000 records of bank credit transaction is shown as Figure 4.

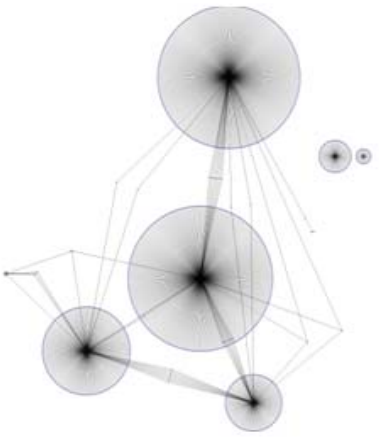


Figure 4. The graph of balloon layout for 1000 records of bank credit transaction

E. The Net Layout Algorithm

The net layout applies the spring-embedded model presented by Peter Eades [11]. The spring-embedded model is a more mature layout algorithm, which has give birth to a number of force-directed layouts. Its principle is: take each node as a steel ring, and the edge was seen as a spring connecting the various steel rings. At first node can be distributed in any location and calculate the force acting on the ring. The algorithm is a cyclic process declining the energy state of the physical system ceaselessly and reaches equilibrium eventually. The view from the geometry means that nodes and edges will be distributed uniformly without the need to determine the distribution of each node in graphic, because a physical system can have multiple equilibriums, which is the biggest advantage of this algorithm.

Based on the skeleton sub graph theory, the key idea of improved net layout is to decompose the original graph into a skeleton sub graph and several stub trees, and to layout them with force-directed layout algorithm. The experiments and analysis indicate that our algorithm outperforms the traditional K-K [12] algorithm when the size of the graph is smaller than a certain constant, and the result seems to be easier to lead the user to identify the skeleton sub graph and the stub trees, and to understand the original graph.

Skeleton sub graph is a special sub graph defined in the power-law characteristic graph, which can be got by filtering suspension point to the original graph repeatedly, but also to identify multiple stub trees attached to the skeleton sub graph. The skeleton sub graph is a collection of nodes that degree relatively high and quantity relatively small, while the stub tree opposite. The power-law characteristic guarantees the non-uniform feature of node degree distribution effectively. In contrast, in order to achieve a better layout result, the time spending of stub tree layout algorithm mainly on the process of selecting layout area.

The net layout is suitable for large-scale data analysis, especially for relational databases. It can analysis a variety of social relationships accurately, which revealed its deep structure, and it also explains a series of contemporary social phenomena in-depth and concrete, while avoiding the overlap issue between the entities. Typical applications can be found such as kinship among people, relation between purchase and sale in circulation and so on. The graph of net layout for 10000 records of purchase and sale is shown as Figure 5.

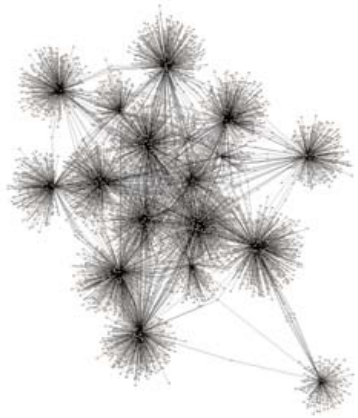


Figure 5. The graph of net layout for 10000 records of purchase and sale

IV. CONCLUSIONS

In this paper, we have discussed about five different kinds of layout algorithms for visualization. We analyzed the features of each layout algorithm and present some suggestions to users on how to choose layout algorithm for specific areas. Besides, we have given some necessary improvements to supply a gap for each algorithm. The research in this paper gives an overall analysis of the present visualization techniques, which is vital to the future development of related algorithms and technologies.

ACKNOWLEDGMENT

This project is supported by the Planning Project of the National Eleventh-Five Science and Technology (2007BAK34B04) and the Chinese National Natural Science Fund (60704042) and Aeronautical Science Foundation (20080768004) and the Program of 211 Innovation Engineering on Information in Xiamen University (2009-2011).

REFERENCES

[1] Bruce H. McCormick, Thomas A. DeFanti and Maxine D. Brown (eds.). Visualization in Scientific Computing. ACM Press, 1987.

[2] Battista G D , Eades P , Tamassia R ,et al. Graph Drawing: Algorithms for the Visualization of Graphs [M] . Prentice-Hall, 1999。

[3] Ellson J , Gansner E R , Kout sofios E , et al . Graphviz and Dynagraph2Static and Dynamic Graph Drawing Tools [M] // Junger M ,Mutzel P ,eds. Graph Drawing Software. Springer-Verlag, 2003 :1272148.

[4] BATTISTA G D, TAMASSIA R, TOLL IS I G, et al. Algorithms for the visualization of graphs[M]. [S.l.]: Prentice-Hall, 1999.

[5] M. Baur and U. Brandes, “Crossing Reduction in Circular Layouts”, Proc. Graph-Theoretic Concepts in Computer-Science (WG ’04), 2004, pp.332-343.

[6] U. Do˘grus¨oz, B. Madden and P. Madden, “Circular layout in the Graph Layout Toolkit”, Proc. Graph Drawing (GD ’96), 1996, pp.92–100.

[7] Reingold E.M. and Tilford J.S. Tidier Drawing of Trees. IEEE Transactions on Software Engineering, vol.7, NO.2, 1981, pp.223 – 228.

[8] Walker J.Q. A Node-Positioning Algorithm for General Trees. Software: Practice and Experience, vol.20, NO.7, 1990, pp.685 – 705.

[9] Wills G J . Nicheworks2Interactive Visualization of Very Large Graphs[C] // Proc of Graph Drawing ’97 ,1997 :4032414.

[10] Chun-Cheng Lin, Hsu-Chun Yen. Journal of Graph Algorithms and Applications. vol.11, NO.2, 2007, pp. 431–452.

[11] P. Eades. A heuristic for graph drawing. Congress Numerantium, vol.42, 1984, pp. 149–160.

[12] KAMADA T, KAWA I S. An algorithm for drawing general undirected graphs[J]. Information Processing Letters, vol.31, NO.1, 1989, pp.7 - 15.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 373-376

The Design and Implementation of Ultra-wideband Microwave Amplifier

Hui Xu, and Hongzhuan Feng
School of Information Science & Engineering, Shenyang University of Technology, Shenyang 110870, China
Email: xhimage@163.com, linxi1401@126.com

***Abstract*—Equivalent gain amplification and broadband networks matching of weak microwave signal are difficult in Ultra-wideband microwave receiver system. A highly integrated amplifier IC ADL5541 from ADI is adopted as the core unit in this paper, coordinate with an external matching network. The design can realize an Ultra-wideband microwave amplifier with 15dB gain which works in the 500~4000MHz with +25dBm of high input linearity magnification. S-parameter curves of the amplifier are simulated by calling the S2P file of ADL5541 in Agilent's ADS simulation software, and then the key parameters of external circuits are determined. The experiments show this kind of design can meet the general applications of microwave receiver systems by measurement under the condition of frequency range 500 ~ 2000MHz.**

***Index Terms*—Ultra-wideband, microwave receiver systems, microwave amplifier, highly integrated amplifier, broad-band networks matching**

I. INTRODUCTION

In the microwave receiving system, the first-class in general need is microwave amplifier to amplify weak signals, and then the received signal is mixed with the local oscillator signal to obtain the intermediate frequency (IF) signal component that contains the signal. In the microwave signal launch from the antenna, also needs to be amplified to enough power [1]. In the microwave receiving systems and launch systems, all need to adopt microwave amplifiers [2].

The realization of microwave amplifier uses either discrete components, or the integrated circuit(IC). If the discrete components are adopted to the design, appropriate amplification transistors are needed to select. The impedance matching is worthy of concern when making use of discrete components, and another problem is the power gain of the transistors slow down fast along with the frequency increases at high frequency part [3]. In the practical design of microwave amplifiers, it is often the way to multi-stage cascade effect of broadband to work, each level of solution-oriented as an indicator, then the whole amplifier to secure a satisfactory result [4][5][6]. It can be seen that using discrete pieces building a microwave amplifier cumbersome and difficult to debug. If you select the integrated circuit, can be a good solution to this problem. Because the integrated circuits to impedance matching, bias set up these key points of the design of microwave amplifiers have been designed to be completed, very good package to a complete chip, and the work can be done fairly wide frequency band.

The project is supported by the Department of Liaoning Education. No.2009T074.

II. GAIN BLOCKS ADL5541

The ADL5541 is a broadband 15dB linear amplifier that operates at frequencies up to 6GHz. The device can be used in a wide variety of CATV, cellular, and instrumentation equipment.

The ADL5541 provides a gain of 15dB, which is stable over frequency, temperature, power supply, and from device to device. The device is internally matched to 50Ωwith an input return loss of 10 dB or better up to 6GHz. Only input/output ac coupling capacitors, power supply decoupling capacitors, and an external inductor are required for operation [7]. The ADL5541 consumes 90mA on a single 5V supply

The functional block diagram of ADL5541 shows in Fig. 1.

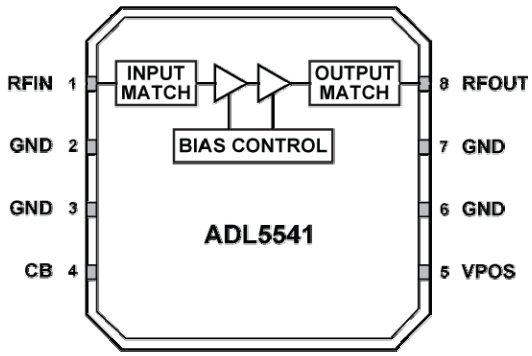


Figure 1. Functional block diagram of ADL5541

ADL5541 is a highly integrated amplifier; in it there integrated input matching circuits and output matching circuits, and bias control circuitry. The typical circuit on instructions of the chip are referenced in actually use, a desired function can be achieved.

III. GAIN BLOCKS ADL5541 SIMULATION IN ADS

The scattering parameter curves is obtained by Agilent's ADS software simulate of ADL5541.Because there is no ADL5541 model in the original library of ADS, the model should be built by user before simulation. The following steps describe how to build simulation models of ADL5541 by the datasheet.

Because the ADS can be used in SnP files, it can be through the establishment of the device SnP file to get the device simulation model. SnP document is called a linear S parameter file (also can be G, H, Y, and Z parameters), n value is a number between 1~99. When n=2, i.e. two-port device, you can include noise parameters. There are two ways building SnP files, one is to manually edit the text files themselves; the other one is from the available data (which may be ADS simulation data obtained from network analyzer can also be obtained and other equipment with test data) to export generation.

A.Manually edit

The S parameter provided by the manufacturer's datasheet is written in prescribed form in Notepad, Save the file as a SnP, then use the Data Items in the SnP module reference to the file, you can put the module as an S-parameter models for circuit-level, system-level simulation. Formats are as follows:

! At the beginning of the statement as a comment prefix statement, these statements in the program is running will be ignored.

At the beginning of the file statement to illustrate the definition of the format, content, including the frequency units, parameter type, data format and normalized resistance value, with a leading # sign to do:

To S2P model as an example to make a specific presentation format is as follows:

freq_units parameter format Rn

<data line>

…

<data line>

Meaning is as follows:

Compiler, followed by the symbol prompts on the parameters

freq_units set up units, parameters are: GHz, MHz, KHz, or Hz

parameter set the parameters, S1P device can be set to S, Y, Z parameters

S2P device can be set to S, Y, Z, H parameters

S3P and S4P parameters can be set to S

Format Content Format DB for db-angle

MA for magnitude-angle

RI for real-imaginary

Rn impedance setting, usually 50 ohms

If the file does not begin with "#" marks the beginning of the option, then the default options are:

GHz S MA R 50

Note: S2P in a "!" for comment.

B. Export data generation

With the HP-IB control, "write" command, you can generate the SnP files.

According to the above manual editing steps to compile ADL5541 the S2P file, and save the file as. S2P. Then, the Project into the ADS environment, the schematic interface, select the Data Items section of the 2port - S parameters file, simulation and analysis. Fig. 2 shows, for the ADL5541 in the ADS simulation analysis diagram.

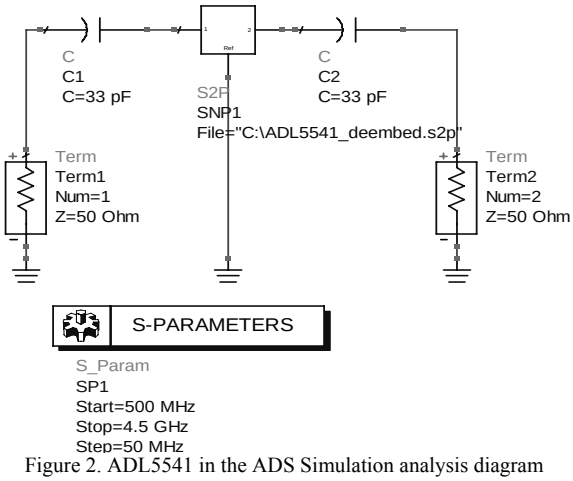


Figure 2. ADL5541 in the ADS Simulation analysis diagram

C1 and C2 in Fig. 2 are AC-coupling capacitors. Simulation results of ADL5541 are shown in Fig. 3.

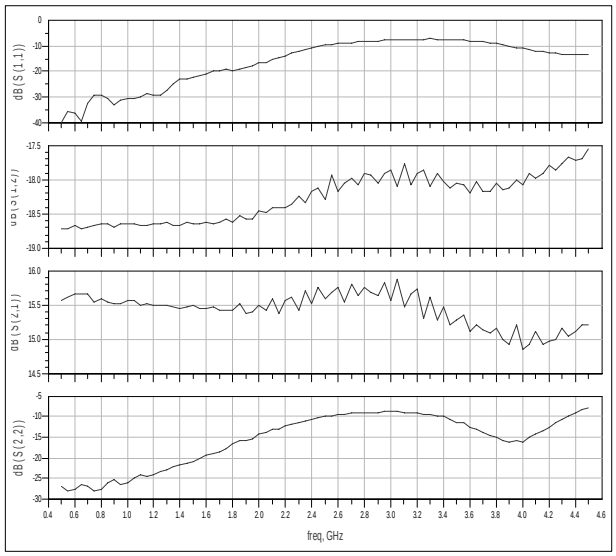


Figure 3. ADL5541 in the ADS schematic of the simulation results

IV. Design of 15dB Gain Amplifier

According to provided chip manuals, reference system design specifications, and determine the 15dB gain amplifier schematic diagram shown in Fig. 4.

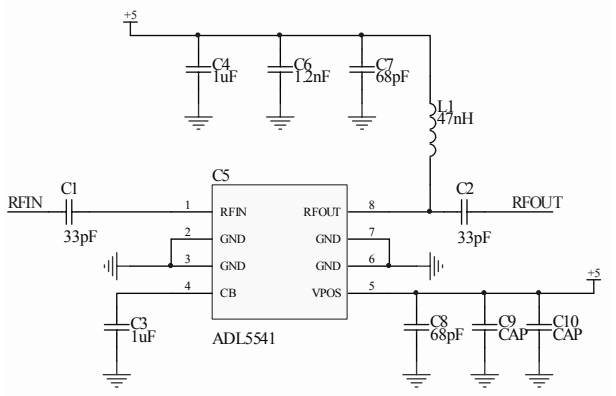


Figure 4. 15dB gain amplifier schematic diagram

In the Fig. 4, C1 and C2 are same as in Fig. 2, C1 is used to cut off the DC component contained in input RF

signals to prevent the DC component of input signal from affecting the amp. And C2 is output AC coupling capacitor to prevent high DC component of ADL5541's output from entering the following circuit to cause dangerous, for the DC voltage of ADL5541 port 8 pin is up to 5V.

The debugging of simulation and experiment through ADS show that the chip work in a high frequency, in accordance with instructions chips prompted, select the patch package 33pF ceramic capacitors can be. Port 4 is a low-frequency channel, to pick up a large capacitor to ground, select the recommended value of 1μF capacitor. The parallel capacitor of VPOS C8, C9, and C10, in which C8 can select the recommended value 68pF, or select 62pF of ceramic capacitors. C9, C10, mainly in the use of debugging. When the power signal is mixed with noise, you can select larger capacitors, and the components are placed to follow this rule: C8 is closest to the chip, C8, C9, C10 are close to each other as near as possible. C4, C6, C7 are decoupling capacitors for the power supply, because the general linear RF circuit chip is very sensitive to the power supply noise and it is necessary to eliminate all of the power supply noise by these capacitors. C4 is determined by the measured value of 1μF, C6 is 1.2nF, C7 can use a value of 68pF or 62pF.

The chip output RFOUT is connected to inductor L1 which constitutes a DC bias circuit, L1 uses a value of 47nH.

V. Implement of 15dB Gain Amplifier

The selection of media of PCB plate is more critical during RF circuit design. The plate with larger dielectric constant and thin high-frequency plate is better choice in general. It is discovered that the width of microstrip line is wider with smaller dielectric constant and thicker thickness of PCB through computation. The microstrip line width probably is 5mm for ordinary PCB, and the microstrip line width decrease to 1.89mm when high-frequency plate FR-4 is used with which dielectric constant is as high as 4.4, the plate thickness is as thin as 1mm. For the RF circuit, the components placing, pad setting and chip welding brings great benefits. When actual implement, the PCB plate selects the dielectric constant of 4.1 common plates, the RF signal lines are using microstrip circuit design. With rigorous calculation, microstrip line impedance matching to 50Ω. Both sides microwave signal transmission microstrip line should be evenly distributed ground vias, and via diameter is 0.4~0.5mm, This is to ensure that the shortest path to ground to achieve a shielding effect, and also can be a very good radiation. When the power line designing, width as wide as possible, so that you can reduce the size of the loop resistance. All the wiring far away from the PCB board about 2mm, it will prevent the wires contact the shielded box body. In the PCB board, area where without signal lines should be spreaded with unified copper shop, the PCB board's bottom surface as well, and circuit are grounded.

Finally, 15dB gain amplifier, the physical map in Fig. 5.

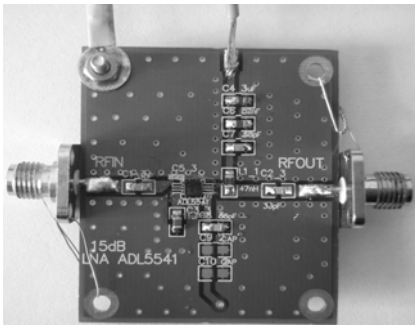


Figure 5. Physical map 15dB gain amplifier

VI. The Experimental Verification of Microwave Amplifiers

The existing equipment in Lab are Agilent8648B RF signal generator (frequency range: 9kHz~2000MHz, power range: −136dBm~14.5dBm), Agilent U2001A USB Average Power Sensor (input frequency range: 10MHz~6GHz, power detection range: −60dBm~20dBm), Anritsu MS2665C spectrum analyzer, the standard +5V DC power supply, as well as a number of coaxial cables.

The following experiments are carried out on the condition below frequency 2000MHz because the restrict of existing signal generator.

A. Experimental steps

First, two coaxial cables used for this experiment are connected between the signal generator and average power sensors respectively to measure coaxial cable attenuation. The output power of RF signal generator is −15dBm, beginning frequency is 500MHz, frequency step is 50MHz, and end frequency is 2000MHz. The attenuation of coaxial cable over the measuring frequency range is recorded. Then, the 15dB gain amplifier circuit is connected to the measurement system through the two coaxial cables, and the gain of signal over the measuring frequency range is obtained. Another experiment is carried out by replacing the average power sensor by power spectrum analyzer for frequency offset measurement.

B. Experimental results

In accordance with the requirements of experimental procedures, we record experimental results, and obtain the power output vs frequency plot in the Matlab. The measurement result by average power sensor is shown in Fig. 6, and the measurement result using spectrum analyzer is shown in Fig. 7.

From Fig. 6 and Fig. 7, we can see that there is a big loss of coaxial cable and its loss rule is not along with the frequency increases, but nonlinear variation. Observing the figures at the top position of the curve, they indicate that add amplifier before and add amplifier later, the difference of the signal output power, it represent the magnifying power ability of the gain amplifier. By analysis, it is discovered that both approximately in 15dBm place, therefore, it indicates that the 15dB gain amplifier basically satisfies the design requirements.

Considered using spectrum analyzer measuring signal's power is not very accurate generally, therefore, when carrying out power measurement, is not recommended to use the spectrum analyzer. In the experiment, we adopt spectrum analyzer's primary reason is that using it observe the RF signal, which enlarged after the gain amplifier whether to have the frequency offset phenomenon and whether to introduce the new frequency.

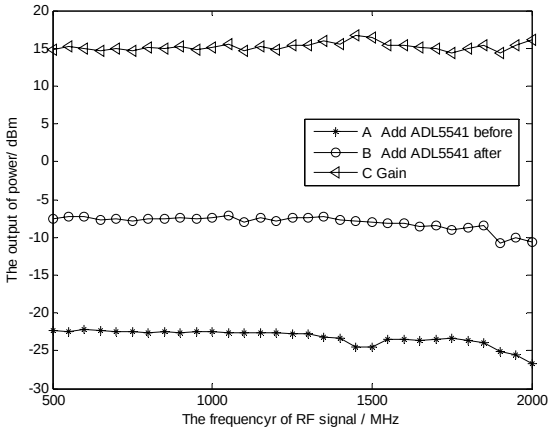


Figure 6. Results is that using average power sensor

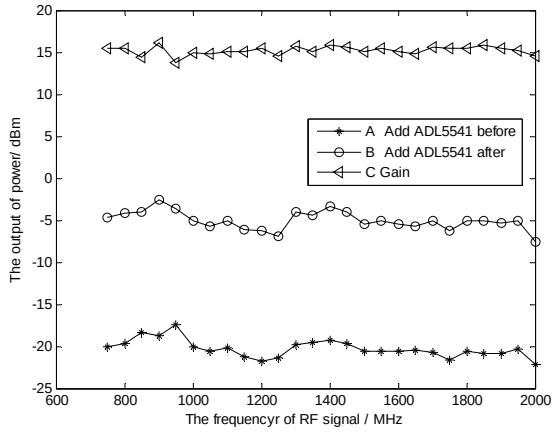


Figure 7. Result is that using spectrum analyzer

C. 15dB gain amplifier debug points

In the process of 15dB gain amplifier debugging, we primarily examine its input port voltage and the port 4 voltage. If the input voltage about 2.8V, the port 4 voltage about 3.2V, it represents amplifier to work. And too large or too small, it cannot achieve the enlargement function.

VII. CONCLUSION

The present paper designs the microwave amplifier which works in UWB (ultra-wide band). By experiment, it is discovered that the gain block ADL5541 input AC coupling capacitor and the output AC coupling capacitor for constructing the filter network to play the crucial role for the gain block ADL5541.

When the input signal for low frequency, the range is 50MHz~500MHz, input AC coupling capacitor and the output AC coupling capacitor value should be large, generally takes 0.1μF; When the operating frequency exceeding 500MHz, selects the 33pF SMD package capacitor.

This paper represents that by measuring average power, the microwave amplifier to work in the frequency range up to 500~4000MHz, input linearity of +25dBm, the gain to 15dB. Meet the general applications of microwave receiver systems.

ACKNOWLEDGMENT

The project is supported by the Department of Liaoning Education. No.2009T074.

REFERENCES

[1] M. M. Radmanesh, *Radio Frequency and Microwave Electronics Illustrated.* Prentice Hall PTR, 2005.

[2] D. M. Pozar, *Microwave Engineering (Third Edition)*. John Wiley & Sons Inc, 2005.

[3] X. F. Liu, D. H. Wang, S. D. Wang, "0.8~8.5GHz Broadband Low Noise MMIC Amplifier", *Semiconductor Technology, China,* Vol.133, No.16, pp.514-519, June 2008.

[4] K. W. Qian, Z. Tian, "Development of ultra-broad band LNA in 0.1~2.8GHz", *ELECTRONIC COMPONENTS AND MATERIALS, China,* Vol.27, No.8, pp.62-64, Aug 2008.

[5] C. M. Chen, "Design for 1~8GHz broad-band microwave power amplifier", *ELECTRONIC MEASUREMENT TECHNOLOGY, China*, Vol.31, No.2, pp.73-86, February 2008.

[6] G. D. Ge, H. Y. Kang, S. M. Li, "Design of 0.35-2.5GHz ultra-broad band low noise amplifier", *Microcomputer Information*, Vol.24, No.2, pp.275-276, 2008

[7] ADI Inc, *50MHz to 6GHz RF/IF Gain Block ADL5541*. www.analog.com, 2007.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 377-380

Rate Adaptation Transcoding Control Algorithm for Video Transmission over Wireless Channels

Wenbing Fan [1], Minglin Zhou [2], and Yingqiao Shi[1]
[1] School of Information Engineering, Zhengzhou University, Zhengzhou, China
Email: iewbfan@zzu.edu.cn
[2] Xinyang Agricultural College, Henan, China

***Abstract*—In this paper, we investigate the problem of rate adaptation transcoding algorithm for transmitting video over wireless channels, i.e., channels such that errors tend to occur in clusters during fading periods. An adaptive rate control algorithm based on the theory of stochastic optimal control was proposed, which is capable of dynamically determining the transcoder's objective bit rate according to the bandwidth variation of wireless channel and the buffer occupancy. In addition, it can balance real-time transmission with continuity of video playing and attempt to acquire the best overall performance. Then we analyze the transient performance, steady performance, and computational complexity of the algorithm. Our experimental results demonstrate that the proposed algorithm can accurately control the bit-rate of the transcoded video stream and reduce the number of frames been skipped without violating the end-to-end delay requirement.**

***Index Terms*—wireless LAN, rate control, adaptation transcoder, real-time transmission**

I. INTRODUCTION

Recently, the increasing demand for mobile communications has resulted in the extensive use of wireless communication technology. Transmission of real-time video over wireless nextworks is challenging because of the delay constraints involed, and because of the negative impact of channel errors on the perceptual quality of video at the decoder. The video stream transmission usually goes through different network segments, e.g., from wired segment to wireless segment. Because of the difference of network properties, a transcoder is needed when the video stream spans different network segments. Specially, in the current wireless video communication systems, the wireless link is usually an extension of the wired network, i.e., the wireless access point is a node of the wired network and mobile terminals connected to the access point through a point-to-point link. The video source encdoer is not generally located right at the wireless access point or base station. In order to adapt the objective bit rate of transcoder to the variation of wireless channel, it is required to buffer the bit stream. However, this will increase the transmission delay, and excessively delay is not tolerable for the applications that have real-time requirement[1].

Many rate control algorithms are evidently standard coders, such as those based on MPEG or H.263, where variable length coding is used or where compression involves a predictive coding scheme[2], but the above algorithms are designed for the wired channel whose bandwidth variation is relatively small, consequently are not suitable for the wireless transcoder[3]. These techniques can guarantee certain error rate requirement for the worst channel condition. However, this causes unnecessary overhead and wastes bandwidth when the channel is in a good state. H.264/AVC is a new generation video compression standard, which is proposed by joint video team (JVT) and its suggested rate control methods include JVT-F086[4], and JVT-G012[5], *etc*.. JVT-F086 is improved on the basis of TM5. JVT-G012 conducts the linear prediction for the mean absolute deviation (MAD) of current macroblocks, then calculates the quantization parameters based on the quadratic rate distortion model used in MPEG4. However, both JVT-F086 and JVT-G012 are also not competent for the wireless channel whose bandwidth variation is more stochastic. In addition, the above rate control algorithms are usually devoted to the macroblock layer bit allocation.

In this paper, we concentrate on how a real-time wireless video system can be supported by a linear quadratic Gaussian (LQG) control scheme and rate adaptation transcoding techniques. To take full advantage of the stochastic optimal control, we propose to feedback mechanism with the transcoding mechanism at the video transcoder, by which one can achieve a result: the rate for the transcoded video is reduced during the periods of poor channel conditions. Furthermore, we dynamically determine the objective bit rate of transcoder, according to the bandwidth variation of wireless channel and the buffer occupancy.

II. WIRELESS VIDEO COMMUNICATION SYSTEM

A. Figures and Tables

In this paper, we consider a wireless real-time video transmission system with a transcoder as illustrated in Fig.1. The transcoder is deployed in the wireless access point through a high bandwidth, low error rate[6], therefore, the transmission between the video source and accedd point is assumed to be error-free. Compared to the fixed bandwidth of wired channel, the wireless channel is more unpredictable, more stochastic, and has high bit error rate[7]. According to the wireless communication theory, the channel usually declines in the transmission, due to the dispersion and reflection of building, mountain

and plant. The current channel model includes Gaussian white noise model, fading channel model.

This paper adopts the Gaussian white noise mode in the communication system. The bandwidth variation of the wireless channel approximates to the Gaussian white noise distribution. Consequently let the effective bandwidth R (kbps) follows the Gaussian distribution with the mean μ (kbps) and the variance σ^2, namely $R \sim N(\mu, \sigma^2)$. Its density function f_R is:

$$f_R = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(R-\mu)^2}{2\sigma^2}} \quad (1)$$

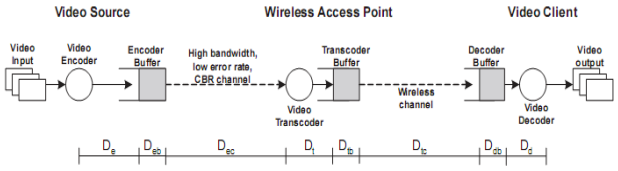


Fig.1. Wireless video communication system

In Fig.1, let us assume that a video is encoded and transmitted at a fixed frame rate F. At the decoder side, the video is decoded and displayed at the same frame rate F. Then, the end-to-end delay each frame experences (from the time it is obtained from the video input to the time it is placed in the video display) consists of several delay components as in Eq.(1).

$$D = D_1 + D_2 + D_3 \quad (2)$$
$$D_1 = D_e + D_t + D_b$$
$$D_2 = D_{ec} + D_{tc}$$
$$D_3 = D_{eb} + D_{tb} + D_{db}$$

Where the D_1, D_2, D_3 stand for the Processiong Delay, Transmission Channel Delay and Buffer Delay, respectively and the subscripts e, t and d stand for the encoder, transcoder and decoder, respectively. Therefore, we can assume the processing delay components are constant and can be neglected. This paper focuses on the relationship between the waiting delay spent in the transcoding buffer and the wireless channel in order to reduce the end-to-end delay through adaptive rate control and optimal synthetic performance.

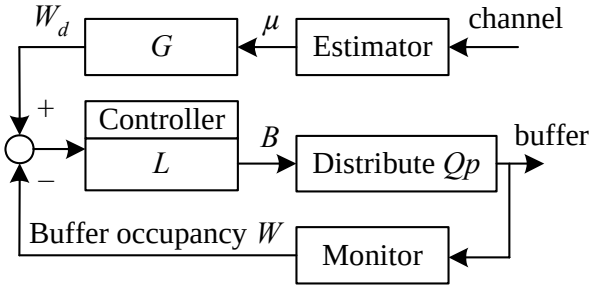


Fig.2 Block diagram of system architecture

III. Adaptive Rate Control Algorithm

A. System Architecture

The block diagram of system architecture is shown in Fig.2. The estimator evaluates the average effective bandwidth of channel and the module *G* determines the desired buffer occupancy (the set point) W_d.

The monitor samples the buffer occupancy of transcoder *W* periodically. The controller *L* determines the objective bit rate *B* according to *W* and W_d. Finally, the algorithm allocates Q_P for macroblocks. The minus in Fig.2 means the negative feedback. If $W > W_d$, the controller decreases the number of bits of frame, else increases its number of bits. Video stream is encdoed according to the specific value and output channel buffer.

B. Adaptation Rate Control Algorithm

According to TMN8 testing model, the variation of buffer occupancy can be described by the following equation:

$$W_{k+1} = W_k + B_k - R_k/F, k = 1, 2, \ldots \quad (3)$$

where W_k is the buffer occupancy in the *k*th sampling point, especially, W_1 is the number of bits pre-buffering. B_k is the number of bits that enters the buffer during the *k*th frame interval. *F* is frame rate. R_k represents the bandwidth of wireless channel in the *k*th sampling point. Note that the following relation exists between D_{tb} and *R*:

$$D_{tb} = \frac{W}{R} \quad (4)$$

It is known that the higher the buffer occupancy is, the longer the end-to-end delay is. On the other hand, the lower the buffer occupancy is, the higher probability of empty buffer is, and the more serious the jitter is. Therefore, this paper introduces a desired buffering delay of transcoding D_d to solve this problem. D_d is a compromise between the end-to-end delay and the playout quality, which can be tolerated by the audience.

In order to realize the rate control described in Fig.2, we transform D_d to a desired buffer occupancy W_d. From Eq. (1),

$$W_d = \mu D_d / 1000 \quad (5)$$

where the μ (kbps)denotes average bandwidth of wireless channel, D_d (ms) denotes desired transcoding buffer delay.

The variation of buffer occupancy is related to the size of current frame B_k. When the occupancy is larger than W_d, it is needed to decrease B_k, and contrarily it is needed to increase B_k.

In order to make a compromise between the control objective and the control performance, a quadratic objective function is designed:

$$\min J = E\left\{\sum_k \left[a(W_k - W_d)^2 + b(B_k - \frac{\mu}{F})^2 \right]\right\} \quad (6)$$

In Eq. (6), the first quadratic term is the error between W_k and W_d. It is obvious that the error should be kept as

small as possible. The second quadratic term reflects the variation of frame size. The less this term is, the more the objective bit rate matches with the channel rate, and consequently the more efficient the wireless channel is. Here a, b>0, both a and b are two weight coefficients, whose values represents the relative importance between two quadratic metrics. The symbol E represents math expection.

Because distribution characteristic of wireless channel is invariable within some time window, Eq. (3) is changed into:

$$W_{k+1} = W_k + B_k - \mu/F - (R_k - \mu)/F,\ k = 1, 2, \ldots \quad (7)$$

Let $x_k = W_k - W_d$, $u_k = B_k - \mu/F$, $e_k = -(R_k - \mu)/F$, the above equation is changed into:

$$x_{k+1} = x_k + u_k + e_k \quad (8)$$

Since $R_k \sim N(\mu, \sigma^2)$ and e_k follows the Gaussian distribution with mean 0 and variance σ^2 / F^2, namely, $e_k \sim N(0, \sigma^2 / F^2)$. Thus Eq. (6) is changed into:

$$\min J' = E\left\{\sum_k a x_k^2 + b u_k^2\right\} \quad (9)$$

Obviously, when $x_k = 0$, the buffer occupancy is equal to W_d. According to the LQG theory, to make the J minimum, u_k should satisfy the following linear condition:

$$u_k = -Lx_k \quad (10)$$

Where the minus means the negative feedback, the controlling parameter L is a positive constant. According to the determinacy and equivalence principle[13,15], the condition of solving LQG problem is the same with that of solving determinacy linear quadratic (LQ) problem. Thus, the optimal L, denoted L^*, can be obtained by the discrete Riccati equation[15,16]:

$$L^* = -\frac{1}{2}\left(\frac{a}{b} - \sqrt{\left(\frac{a}{b}\right)^2 + 4\frac{a}{b}}\right) \quad (11)$$

Based on L^*, Eq. (8), and the definition of x_k and u_k, the number of bits of current frame is:

$$B_k = L^*(W_d - W_k) + \mu/F \quad (12)$$

Thus, the frame layer rate control algorithm on the transcoder (Eqs. (11) and (12)) and its parameters can be obtained which is shown as Fig.3.

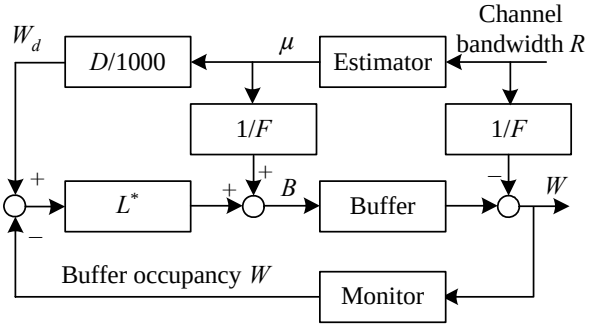


Fig.3 Rate control algorithm of transcoder

IV. Experiment and Results

In order to show the effectiveness of the proposed rate adaptation transcoding algorithms, we implemented several experiments and performed simulations. In the experiments, we use the TCP/IP protocol in the transmission layer. Let a = 0.5, b = 0.5, consequently L^* = 0.618; Z = 0.5, thus the threshold for frame skipping is twice as large as the buffer set point. Let D_d = 30ms and frame rate be 24F/s.

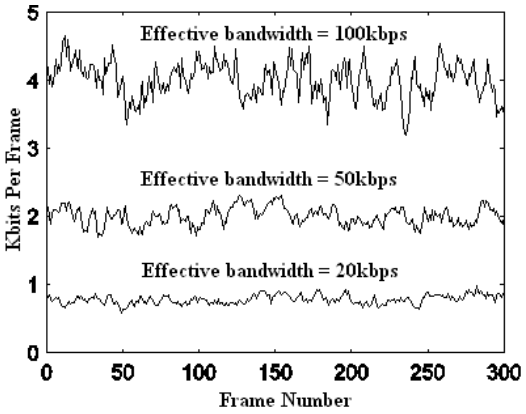


Fig.4 Number of bits per frame for different bandwidth

Fig.4 shows the number of bits per frame on the condition that the initial buffer is equal to W_d when the average bandwidths are respectively 20kbps, 50kbps, and 100kbps.

Fig.5 shows the variation of buffer occupancy on the condition that the initial buffer equals to W_d when the average bandwidths are respectively 20kbps, 50kbps, and 100kbps. It is seen that the average bandwidth is, the W_d is, consequently the larger the variation scope of buffer occupancy is. In practice, the frame size in the frame sequence is variable, thus the bigger the average bandwidth is (e.g. 100bps), the more intense frame size variation it can tolerate, which leads to more fluctuation of buffer occupancy. This phenomenon indicates that our algorithm is more flexible and is capable of choosing different adjustment policies according to different channel statuses.

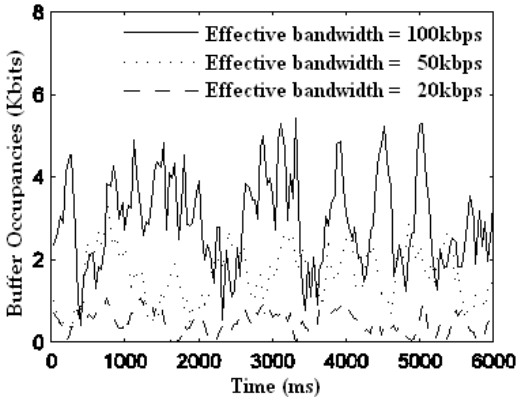


Fig.5 Buffer occupancies for different average bandwidth

On the other hand, when the bandwidth is low, the algorithm decreases the delay by selecting occupancy. However this will lead to a higher probability of empty buffer.

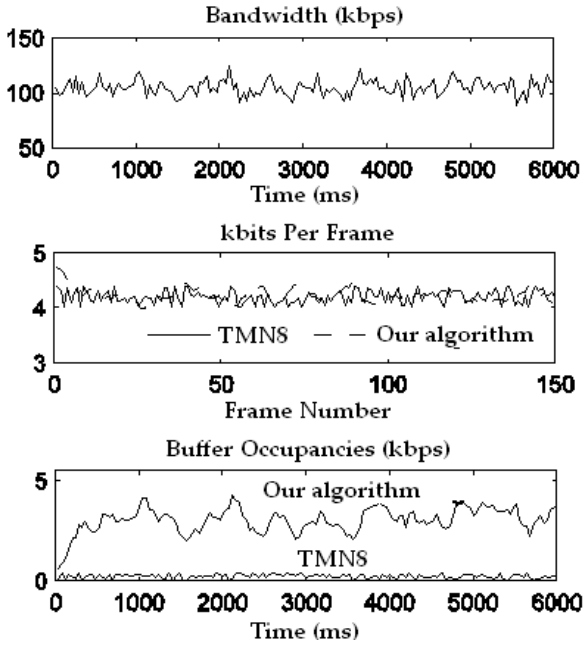


Fig.6 Compared with TMN8

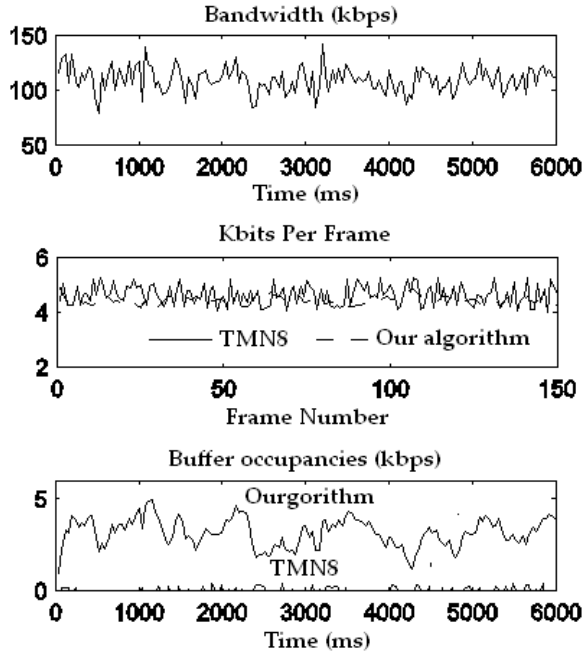


Fig.7 Compared with TMN8

Fig.6 and Fig.7 are the results achieved respectively from our algorithm and TMN8. The average bandwidth is 100kbps. It can be seen that when the bandwidth variation is not obvious (Fig.6), our algorithm and TMN8 have the similar results. In comparison, when the bandwidth varies dramatically (Fig.7), the performance of TMN8 decreases rapidly and the probability of empty buffer occurs, whereas our algorithm still keeps an encouraging performance. This is because our algorithm adopts the new adaptation mechanism.

V. CONCLUSIONS

In this paper, we propose to use a video rate adaptation transcoder in real-time wireless vodeo system. By analyzing buffer delay of video transcoder and making a compromise between the control objective and the control performance, a quadratic objective function is designed. A rate control algorithm is suitable for the transcoder over wireless channel, and designs a frame layer bit rate allocation policy. The experiment results show that our algorithm is capable of adapting the bit rate to the wireless bandwidth variation, and achieving an optimal compromise between the end-to-end delay and the jitter to fully utilize the wireless channel. Therefore, the algorithm improves the synthetic performance of rate control.

REFERENCES

[1] Sadka A H. Compressed video communication. New York: John Wiley & Sons, 2002.

[2] C.-Y. Hsu, A.Ortega, M.Khansari. Rate control for robust video transmission over burst-error wireless channels. IEEE Joural on Selected Areas in Communications, 17(5),1999,756-773.

[3] Lei Z, Georganas N D. A rate adaptation transcoding scheme for real-time video transmission over wireless channels. Singal Processing: Image Communication, 2003, (18):641~658.

[4] Ma S, Gao W, Y Lu, et al. Proposed draft description of rate control on JVT standard. JVT-document JVT-F086, Awaji, Japan, Joint Video Team of ISO/IEC JTC1/SC29/WG11 and ITUT SG16/Q.6, Dec. 2002.

[5] Li Z G, Pan F, Lim K P. Adaptive basic unit layer rate control for JVT, JVT-G012, 7th meeting[C], Pattaya, Thailand, 2003.

[6] Zhijun L, Nicolas D. A rate adaptation transcoding scheme for real-time video transmission over wireless channels. Signal Processing Image Communication, 18(2003)641-658.

[7] Baldi M, Ofek Y. End-to-end delay analysis of videoconferencing over packet-switched networks. IEEE/ACM Transactions on Net- working, 2000, 8(4):479~492.

[8] Chen X.D, Zhu Q.Z. Adaptive rate control on wireless transcoder. Journal of System Engineering and Electronics, 2007, 18(3):633-640.

.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 381-384

Automatic Detection of Vehicle Activities Based on Particle Filter Tracking

Han Huang[1], Zhaoquan Cai[2], Shixu Shi[3], Xianheng Ma[1], and Yifan Zhu[1]
[1] School of Software Engineering, South China University of Technology, Guangzhou , P.R. China
Email: hhan@scut.edu.cn, bssthh@163.com
[2] Network Center, Huizhou University, Huizhou, P.R. China
[3] Guangdong Xinyue, Communication Investment Corporation Limited, Guangzhou P.R. China

***Abstract*—Automatic detection of vehicle activity by computer vision is one of the most important fields of intelligent video surveillance. This paper proposes an approach of automatic detection for visual vehicle activities based on the traffic video. According to the results of particle filter tracking, the geometry information of vehicle movement is abstracted to understand the behavior of vehicle. There are three components of the based particle filter including particle prediction, sampling updating and Bhattacharyya parameter setting. There are three proposed detection models of vehicle activities including breaking, changing-lane driving and opposite-direction driving which are crucial vehicle activities of common traffic accidents. Finally, several traffic videos are tested, and the results indicate the proposed method and model are efficient, high-detection-rate and robust.**

***Index Terms*—Computer Vision; Image Processing; Particle Filter; Video Tracking; Vehicle Activities Understanding**

I. INTRODUCTION

Artificial detection cannot satisfy the real-time demand, so the research of automatic detection becomes more and more important [1]. Complete visual vehicle activity detection generally includes four consecutive steps: Motion object segmentation, object classification, object tracking and behavior understanding. The topic of our proposed work belongs to visual behavior understanding, recently, considerable attentions have been paid to this research direction.

F. Bremond [2] and Y. Ivanov [3] proposed and adopted a two-step approach to the problem of video understanding: A lower-level image processing visual module is used to extract visual cues and primitive events. This collected information is used in a higher-level artificial intelligence module for the detection of more complex and abstract behavior patterns. Behavior pattern learning and understanding may be thought of as the classification of time varying feature data, i.e., matching an unknown test sequence with a group of labeled reference sequences representing typical or learned behaviors [4].

This work has been supported by Natural Science Foundation of Guangdong Province (9151600301000001), Key Technology Research and Development Programs of Guangdong Province (2009B010800026) and Huizhou City (2009G024), Natural Science Research Project of Huizhou University (C209.0404), and Student Research Project of South China University of Technology in 2009.

AP-PROC-CS-09CN005

Rama Chellappa et al. [5] give a framework for activity inference, they start by computing trajectories from a video sequence, then fit models to these trajectories, and finally compute the similarity between model parameters for inferring about the activity. Xiang et al. [6] present a unified bottom-up and top-down automatic model selection based approach for modeling complex activities of multiple objects in cluttered scenes. It is a challengeable task especially under complicated traffic scenes. By behaviors understanding many traffic accidents can be quickly found, even be prevented [7]. Fernyhough et al. [8] establish the spatio-temporal region by learning results of tracking objects in an image sequence, and construct a qualitative behavior model by qualitative reasoning and statistical analysis.

Other research results of visual behavior understanding focus on human object [9-10] but fewer on vehicle objects.

II. PARTICLE FILTER TRACKING BEFORE MODELING

A. Complete Framework

There are three technology steps including vehicle object segmentation, vehicle object classification and vehicle object tracking before the modeling of vehicle activities.Furthermore, behavior understanding is the key step, which is used to analyze the behavior of the visual vehicle objects to determine whether an activity happens. Motion object segmentation and object tracking are normal steps that can be implemented by available technology tools like OpenCV [11]. For object classification, we only consider vehicle object without any classification method.

A Hybrid subtraction strategy [12] is used to do motion object segmentation. Based on the vehicle object segmentation, a mathematical model of vehicle object is for particle filter tracking [13]. Supposed there are N blob signed as a rectangle with the center (x, y), the height $a = y_{max} - y_{min}$ and the width $b = x_{max} - x_{min}$.

Signed a particle as $P(x, y, a, b, v_x, v_y, w)$, where x, y, a, b are the same item of the vehicle blob, v_x, v_y are the velocities in horizontal axis and vertical axis. w is the weight of the particle in filter. There are three components of the proposed particle filter including

particle prediction, sampling updating and Bhattacharyya parameter setting.

B. Particle Prediction

The function of particle prediction is to confirm the status of particles in the next frame so that the vehicle object can be tracked in the following frames. The process of particle prediction can be described as follows.

A vehicle blob can be signed as $C(x, y, a, b)$. There are N particles $P_1, ..., P_N$.

Step 1: generate a random number r of normal distribution. For $r < 0.5$ and $i = 1, ..., N$ run:

$P_i.x = C.x$, $P_i.y = C.y$, $P_i.a = C.a$, $P_i.b = C.b$

Step 2: For $i = 1, ..., N$ and 5 random parameters $r_1, ..., r_5$, $S = (C.a + C.b) \cdot 0.5$,

$P_i.x = P_i.x + Pos \cdot S \cdot r_1$, $P_i.y = P_i.y + Pos \cdot S \cdot r_2$,

$P_i.v_x = P_i.v_x + 0.1 \cdot Pos \cdot S \cdot r_3$, $P_i.v_y = P_i.v_y + 0.1 \cdot Pos \cdot S \cdot r_4$,

$P_i.a = P_i.a \cdot (1 + Size \cdot r_2)$, $P_i.b = P_i.b \cdot (1 + Size \cdot r_2)$.

Step 3: Output the updated particles $P_1, ..., P_N$.

C. Sample Updating

After the particle prediction, a sample updating process is used to update the status of particles for real tracking.

There N particles $P_1, ..., P_N$ and N new particles $PR_1, ..., PR_N$.

Step 1: $sum = \sum_{i=1}^{N} P_i.w$.

Step 2: For $i = 1, ..., N$, generate a random number r of normal distribution and a threshold $T_i = r \cdot sum$.

Find j if $\sum_{i=1}^{j} P_i.w > T$ and $\sum_{i=1}^{j-1} P_i.w \le T$.

$PR_i.x = P_j.x$, $PR_i.y = P_j.y$, $PR_i.a = P_j.a$, $PR_i.b = P_j.b$, $PR_i.v_x = P_j.v_x$, $PR_i.v_y = P_j.v_y$, $PR_i.w = 1$.

Step 3: For $i = 1, ..., N$, $P_i.x = PR_i.x$

$P_i.y = PR_i.y$ $P_i.a = PR_i.a$ $P_i.b = PR_i.b$

$P_i.v_x = PR_i.v_x$ $P_i.v_y = PR_i.v_y$ $P_i.w = PR_i.w$.

Step 4: Output the updated particles $P_1, ..., P_N$.

D. Bhattacharyya Parameter

Bhattacharyya parameter is used to evaluate the matching rate of the object and the prediction. The particle filter will choose the particle of highest matching rate as the tracking object. (x_i^c, y_i^c) is the i-th pixel of the objective blob, and $N_C = C.a \cdot C.b$. (x_i^c, y_i^c) is the i-th pixel of the tracked blob $A(x, y, a, b)$, and $N_A = A.a \cdot A.b$. For $i = 1, ..., N_C$, run

$q_u(i) = c_1 \cdot f(\| (x_i^C, y_i^C \|^2)) \delta((grey(x_i^C, y_i^C) - T_1))$,

$f(\| (x_i^C, y_i^C \|^2) = (\frac{x_i^C - x_0^C}{C.a})^2 + (\frac{y_i^C - y_0^C}{C.b})^2$ and

$$\delta((grey(x_i^C, y_i^C) - T')) = \begin{cases} 1 & grey(x_i^C, y_i^C) \ge T_1 \\ 0 & grey(x_i^C, y_i^C) < T_1 \end{cases},$$

where (x_0^C, y_0^C) is the center of blob $C(x, y, a, b)$, c_1 is a standard constant of normalization. $grey(x_i^C, y_i^C)$ is the gray value of (x_i^C, y_i^C). T_1 is threshold. For $i = 1, ..., N_A$, run

$p_u(i) = c_2 \cdot f(\| (x_i^A, y_i^A \|^2)) \delta((grey(x_i^A, y_i^A) - T_2))$,

where (x_0^A, y_0^A) is the center of blob $A(x, y, a, b)$, c_2 is a standard constant of normalization. $grey(x_i^A, y_i^A)$ is the gray value of (x_i^A, y_i^A). T_2 is threshold.

Finally, Bhattacharyya parameter can be calculated by

$$\rho = \sum_{i=1}^{N_C} \sqrt{q_u(i) \cdot p_u(i)} \Big/ \sum_{i=1}^{N_C} q_u(i) \cdot \sum_{i=1}^{N_A} p_u(i) \quad (1).$$

E. Process of Particle Filter Tracking

The process of particle filter tracking can be described as follows.

Algorithm: Particle filter tracking

Input: Objective blob $O(x, y, a, b)$

Step 1. Particle prediction

Step 2. Update $P_i.w = e^{(\rho - 1)/(2 \cdot S)}$, where ρ is the Bhattacharyya parameter.

Step 3. Sample updating.

Step 4. Calculate the results of the tracked blob by

$$O.x = \frac{1}{N} \sum_{i=1}^{N} (P_i.x \cdot P_i.w) \quad O.y = \frac{1}{N} \sum_{i=1}^{N} (P_i.y \cdot P_i.w)$$

$$O.a = \frac{1}{N} \sum_{i=1}^{N} (P_i.a \cdot P_i.w) \quad O.b = \frac{1}{N} \sum_{i=1}^{N} (P_i.b \cdot P_i.w)$$

Output: Tracked blob $O(x, y, a, b)$

III. Modelling of Vehicle Activities

There are three considered vehicle activities including breaking, changing-lane driving and opposite-direction driving which are modeled by the plane-geometry information of visual vehicle objects.

A. Model of Breaking

Signed $H(x(t), y(t), a(t), b(t), v_x(t), v_y(t))$ as a vehicle object, $(x(t), y(t))$ is the center of the vehicle object at the t-th frame. $v_x(t), v_y(t)$ are velocities at the t-th frame calculating by $H.v_x(t) = H.x(t) - H.x(t-1)$ and $H.v_y(t) = H.y(t) - H.y(t-1)$.

If u is small and close to zero, the vehicle can be considered to be breaking, where $v'_x(t) = \frac{1}{5}\sum_{i=t-4}^{t} v_x(i)$, $v'_y(t) = \frac{1}{5}\sum_{i=t-4}^{t} v_y(i)$ and $u' = \sqrt{H.v'_x(t) + H.v'_y(t)}$. Furthermore, $ax(t) = H.v'_x(t)/(H.a(t)+H.b(t)) - ax(t-1)$ for $t \geq 1$ and $ay(t) = H.v'_y(t)/(H.a(t)+H.b(t)) - ay(t-1)$. For $t = 0$, $ax(0) = H.v'_x(0)/(H.a(0)+H.b(0))$ and $ay(0) = H.v'_y(0)/(H.a(0)+H.b(0))$.

Thus, the breaking detection condition of vehicle H can be described by $B = (u' < T_3) \wedge (ax(t) < 0 \vee ay(t) < 0)$, where T_3 is a threshold calculated by $T_3 = \alpha \cdot (H.a(t)+H.b(t))$ and α is ratio of the number of the blob pixels to the number of the frame pixels.

B. Model of Opposite-direction Driving

Opposite-direction driving is a serious behavior against the traffic rule, which may causes vehicle crashing and other traffic accidents. In artificial detection, the vehicle is considered to be wrong-direction driving when the vehicle is investigated to be moving in the wrong direction of the video.

If the moving direction of the vehicle blob has an included angle $os(i)$ of 90-180 degree with the right direction, an accident of wrong-direction moving is detected. The included angle can be calculated by the Formula 2.

$$os(i) = \arg\cos\left(X_{avg}(i)/V(i)\right) \quad (2)$$

where $os(i)$ is the included angle of the vehicle object at the i-th frame. $V(i)$ is calculated by Formula 2.

$$V(i) = \sqrt{X_{avg}(i)^2 + Y_{avg}(i)^2} \quad (3)$$

where $X_{avg}(i)$ and $Y_{avg}(i)$ are the velocities on average per m frames, which is calculated by Formula 3-5.

$$X_{avg}(i) = \frac{1}{m}\sum_{j=i-m}^{m} x(j) \quad (4)$$

$$Y_{avg}(i) = \frac{1}{m}\sum_{j=i-m}^{m} y(j) \quad (5)$$

where $x(j)$ and $y(j)$ are produced by the step of vehicle object tracking.

C. Model of Changing-lane Driving

Changing-lane driving is a common activity which may lead to side swipe accident of vehicles.

For example, $(x1, y1)$ is the position of geometry center at the i-th frame, and $(x2, y2)$ is the position at the $i+k$-th frame. Signed (x_0, y_0) as the intersection point of the datum line and the line passing through $(x1, y1)$ and $(x2, y2)$, the detection model of wrong-lane driving activity can be shown by Formula 6, with the conjecture that $x1 < x2$.

$$D_{wd} = (x1 < x_0 < x2) \wedge (y1 < y_0 < y2) \quad (6)$$

The tracked vehicle is considered to be wrong-lane driving if $D_{wd} = 1$ that means that the vehicle moves to a wrong lane during driving on the road.

IV. EXPERIMENT RESULTS

This section will introduce the experimental results of detection of the three considered vehicle incident by the proposed tracking method and detection model. Several traffic videos are tested with the platform of CPU P4 2.4G and 512MB memory.

The following figure is an example of vehicles by particle filter tracking. All of the vehicles are marked and tracked including cars, buses and vans far and near in the video.

Figure 1. Examples of particle filter tracking

Figure 1 is two sectional drawings of vehicle breaking detection. Table 1 is the general testing results.

Table 1. General results of breaking detection testing

Video	Vehicles	Breaking	mistake	Missing	Success
1	4	1	0	0	1
2	6	4	0	0	4
3	2	1	1	0	1
4	4	2	1	0	2
5	4	0	1	0	0

Table 1 shows the accuracy of vehicle breaking detection is 100% though there are three mistakes of detection in which non-breaking vehicle is considered as the breaking one. Such mistake roots in high sensitivity of the detection which is useful for automatic traffic video surveillance.

Figure 2. Testing cases of changing-lane driving.

Figure 2-3 indicate cases of changing-lane driving and opposite-direction driving. Table 2 and Table 3 are the general testing results.

Table 2. General results of changing-lane driving

Video	Vehicle	changing-lane driving	Missing	Success
1	1	1	0	1
2	1	1	0	1
3	1	1	0	1
4	1	1	0	1
5	5	3	0	3
6	4	3	0	3
7	4	2	1	1

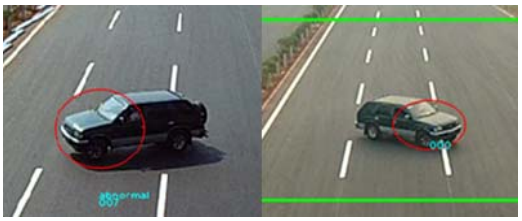


Figure 3. Testing cases of opposite-direction driving.

Table 3. General results of opposite-direction driving

Video	Vehicle	opposite-direction driving	Missing	Success
1	1	1	0	1
2	1	1	0	1
3	1	1	0	1
4	1	1	0	1
5	1	1	0	1

The general results above indicate that the proposed detection model work effectively with few missing. The accuracy if detection is high. Furthermore, the detection rate of the three vehicle activities is 100%. As a result, the proposed model is helpful for the real-time detection of vehicle activities and traffic safety.

V. Conclusion and Disscusion

This paper uses the technology of vehicle object tracking o produce the blob information of vehicle object. Based on the vehicle blob, three detection models are proposed to detect the activities of vehicle breaking, changing-lane driving and opposite-direction driving. The process of particle prediction, sampling updating and Bhattacharyya parameter setting are used to implement the particle filter tracking. Based on the results of tracking, the geometry center and velocity of the vehicle object is used to analyze whether the breaking activities happens. The detection of opposite-direction driving is according to the included angle of the moving direction and right direction. The intersection of lines is the crucial factor for the detection of changing-lane. The experiment of testing 38-vehicle video verifies the high detection accuracy of the proposed model which can be used for real-time automatic detection of vehicle activities. Further study is suggested to extend the tracking method and model to the detection of complex vehicle incidents.

Acknowledgment

The authors would like to thank Mr. Wu Chen, Mr. Zhimin He, Mr. Bo Cao and Mr. Suxin Yao for their programming and initial idea for the proposed model.

References

[1] Y. Liu, P. Payeur, "Vision-Based Detection of Activity for Traffic Control," *Proceedings of IEEE Conference on CCECE 2003 - CCGEI 2003*, Montrbal, May/mai 2003, 2006, 1347-1350.

[2] F. Bremond, M. Thonnat, and M. Zuniga, "Video-understanding framework for automatic behavior recognition," *Behavior Research Methods*, Vol. 30, No. 3, pp. 416-426, 2006.

[3] Y. Ivanov and A. Bobick, "Recognition of Visual Activities and Interactions by Stochastic Parsing," I*EEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, pp. 852-872, 2000.

[4] A. Bobick and J. Davis, "The Recognition of Human Movement Using Temporal Templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 3, pp. 257-267, 2001.

[5] R. Chellappa, A. Roy-Chowdhury, and S. Zhou, "Recognition of humans and their activities using videos," *Morgan & Claypool Publishers*, 2005.

[6] T. Xiang and S. Gong, "Beyond Tracking: Modelling Activity and Understanding Behaviour," *International Journal of Computer Vision*, Vol. 67, No.1, pp:21-51, 2006.

[7] M. Liu, C. Wu, Y. Zhang, "A review of Traffic Visual Tracking technology," Audio *Language and Image Processing, 2008.* ICALIP 2008. International Conference on 7-9 July 2008 Page(s):1016 - 1020

[8] J. Fernyhough, A. G. Cohn, and D. C. Hogg, "Constructing qualitative event models automatically from video input," *Image and Vision Computing*, vol. 18, no. 9, pp. 81–103, 2000.

[9] K. Teddy. "A Survey on Behavior Analysis in Video Surveillance for Homeland Security Applications," *2008. AIPR '08*. 37th IEEE 15-17 Oct. 2008 Page(s):1-8

[10] C. Cohen, F. Morelli, K. Scott, "A Surveillance System for Recognition of Intent within Individuals and Crowds," *IEEE Conference on Technologies for Homeland Security*, Waltham, MA, pp. 559-565, 2008.

[11] Gary Bradski, Adrian Kaehler. Learning Opencv. O'REILLY Press. October, 2008.

[12] Z. Zivkovic. "Improved adaptive Gaussian mixture model for background subtraction," *In: Proceedings of the 17th International Conference on Patter Recognition.* Cambridge, United Kingdom, IEEE, 2004, 2: 28-31

[13] K. Nummiaro, E. Koller-Meier, L. Van Gool, "An Adaptive Color-Based Particle Filter," *Image and Vision Computing* 2003, 21(1): 99-11

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 385-388

K-Multipath Routing Mechanism with Load Balancing in Wireless Sensor Networks

Shaohua Wan[1] , and Yanxiang He [2]
[1] School of Computer, Wuhan University , Wuhan, Hubei 430072, China
shwanhust@gmail.com
[2] School of Computer, Wuhan University , Wuhan, Hubei 430072, China
yxhe@whu.edu.cn

Abstract—**It is desirable to allow packets with the same source and destination to take more than one possible path. This facility can be used to ease congestion and overcome node failures. In this paper, we design and implement a k-multipath routing algorithm that allows a given source node send samples of data to a given sink node in a large scale sensor networks. Multipath routing can increase end-to-end throughput and provide load balancing. However, its advantage is not obvious in wireless sensor networks because the traffic along the multiple paths will interfere with each other. Our multipath routing algorithm tries to keep multipath as node disjoint routes. In order to achieve a minimum mean delay for the whole network, we study the two different policies with which we distribute traffic over different paths. The simulation results reveal that our multipath routing approach does not surprisingly perform better than the shortest path routing (single path routing SPR) in terms of load balancing and quality of data (QoD).**

***Index Terms*- multipath routing protocol; shortest path routing; load balancing**

I. INTRODUCTION

In one of our recent work, we have implemented the split-tree mechanism to prolong the operational lifetime of the nodes, through splitting the root of the tree that can be used concurrently providing many parallel paths from the sub-roots to the sink node for a given query-region, however, yielding better energy consumption than a single path. Motivated by these important results, we feel compelled to study the performance of multipath for each given pair of nodes. When we use the shortest path routing (single path) to ship the aggregated results from the root of the tree to the sink and at the same time, if we need to run the continuous query in excess of 100 hours, the energy of the batteries in the nodes along the shortest route between the root-sink pair will drop faster than their neighboring nodes, leading to undesirable effects as longer delays, congestion increases and lower packet delivery. Finally, those nodes' energy will quickly be depleted, and further shorten the network lifetime. Multipath routing aims to establish multiple paths between source-sink pair of nodes and thus more sensor nodes to be responsible for the routing tasks which in turn, the total traffic are distributed evenly onto multiple routes simultaneously in wireless sensor networks. However, we need to consider the complexity and overhead of the multipath routing. In the case of k-multipath routing mechanism proposed in this paper, maintaining multiple routes to a destination leads to large routing tables at intermediate nodes.

The rest of the paper is organized as follows. In section 2, we provide related work into the area of multipath routing for wireless sensor networks. We model the query component in SIDnet-SWANS [1] and formulate the general construction point-to-point routes problem in section 3. Section 4 discusses route establishment of the k-multipath routing. Based on simulation results, section 5 presents a detailed analysis of load balancing, end-to-end delay and the packet delivery rate metrics for both multipath and the shortest path routing mechanisms. Section 6 concludes the paper.

II. RELATED WORK

We view the routing protocols as two major categories: multipath routing versus single-path routing. In single-path routing, only a single route is used between a pair of source-destination nodes. Two of the most widely used are the Dynamic Source Routing (DSR) [3] and the Ad hoc On-demand Distance Vector (AODV) [4] protocols. AODV and DSR are both on-demand protocols. Also, most of the multipath routing protocols discussed in this paper are an extension of one of these two protocols. [5] presents a multipath routing protocol with a load balancing policy (MRP-LB) to improve the network throughput, decrease average end-to-end delay and reduce congestion, which distributes traffic at packet level [6]. Recently, [2] presents a multi-path computation algorithm to find a set of paths for the given demand and use Integer Linear Programming (ILP) approach to derive an optimal solution to maximize the achievable bandwidth and minimize the required memory size.

III QUERY MODEL AND PROBLEM FORMULATION

A. Qurey Model

Given a goal of allowing users to pose declarative queries over sensor networks, we adopt a SQL-style query syntax. The typical scenario is as follows: a network user connects to one of the sink nodes, formulates and submits a query of the following form:

Q: SELECT ALL/MIN/MAX/AVG (measurement)
FROM Region (R1(x_1, y_1, ... , x_n, y_n))
WHERE Condition (measurement)

AP-PROC-CS-09CN005

FOR Lifetime
SAMPLE EVERY Sampling Interval

R1 represents the geographical bounds of the region in which the samples for the query are to be collected from. Sampling interval indicates the frequency each node must acquire the measurements and ship the data towards the sink. The sensor must stop sensing and sending the data towards the sink node after the lifetime period expires. There are two possible situations based on which the query Q1 is handled at the sink node.

- The sink node is physically located within the sampling region R1
- The sink node is outside of region R1

In our previous work, our approach is designed to exploit the root-load balancing in the scenario where this is readily possible in the situation where the sink node is physically located outside the region R1. Fortunately, this case is more common in large and very large sensor network applications. Figure 1 gives an illustration of this case that we will exclusively consider. Therefore, we will have to construct point-to-point routes from the aggregation root node, which is situated inside the sampling region, to the ultimate destination, the sink node.

B. Problem Formulation

Sensor nodes that are outside the sampling region are also important as they might be used in data-relay duties, making the connection between the producer, in the sampling region, and its consumer, the sink node. For each source-destination pair, a single (shortest) path is always discovered and used for data transmission, as seen in Figure 1, the aggregated information will be sent to the sink through the bold intermediate nodes. Obviously, that area close to the line segment will be very likely to develop hot-spots. The fact that these nodes are overused is one of the major causes for hot spots. This paper provides a new multipath protocol for mitigating the sensor network hot-spot problem, considering load balancing as well as quality of data.

IV. K-MULTIPATH ROUTING ALGORITHM

In our experiment we assume that each node knows its location and the location of its neighbors. This SIDnet-SWANS simulator provides us with the heartbeat algorithm, which is already implemented and executed in the first hour of the simulation, and finds the neighbors for us. The algorithm for constructing the routing structure should be as follows. For a given source-sink pair of nodes, the sink will unicast on a shortest path routing (along a straight imaginary line) the query request to the source node. Subsequently, based on the query specification, referring to Figure 2, we will draw a segment orthogonal to source-sink line segment. We will split the segment in k places, which will correspond to k intermediate destination points (breakpoints) of the paths between the sink and the source. The distance between two consecutive paths on the line which is orthogonal to the source-sink segment will be equal. For each line segment, we forward the data packet using nodes closest to the line segment. Since we have k breakpoints for a given source-sink pair, we will establish k multipath to offer more opportunities for regulating the traffic over the network.

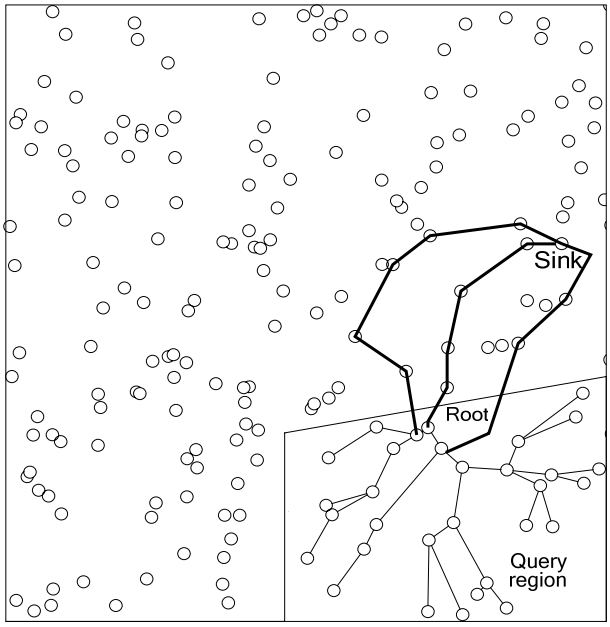


Figure 1. Sink node is physically located outside the sampling region.

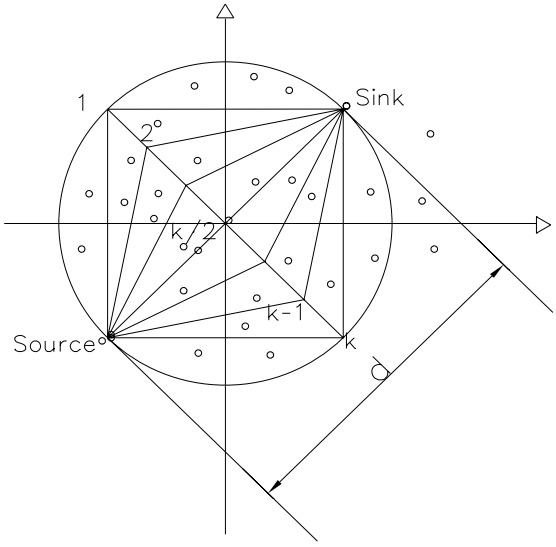


Figure 2. K-multipath routing protocol model.

V. SIMULATION STUDIES

A. Experimental Setup

We use a simulation model based on SIDnet-SWANS. Our simulation setting is as follows. We create 500 nodes uniformly deployed in a $2\times2km^2$ area, which use 802.11 protocol at the MAC layer, and the heartbeat node discovery protocol in order to determine the neighbors. We randomly pick up one pair of nodes as source-sink nodes from the physical terrain, and we don't consider the characteristic of the mobility of the nodes. Although how the number of the paths affects the performance remains unknown, there are 5 paths to be used in the

simulation. We choose a path randomly from the multiple paths with the same probability. Moreover, we try to keep the number of paths odd.

We study two different ways to use the multiple paths. In one method, called multipath routing 1, we choose a path randomly from the multiple paths with the same probability. The other method, called multipath routing 2, is to choose a path with a probability inversely proportional to the length of the path. We vary the sampling rate in order to observe the effects of packet loss in the nodes due to the interferences among the multipath.

B. Performance Evaluation of the Multipath Routing Against That of the Shortest Path Routing

We will compare the performance of the shortest path routing and multipath routing in different aspects. We evaluate the performance according to the following metrics:

1) The load distribution: This metric provides the average relayed traffic in packets as function of the distance to the network center, in accordance with the Pham and Perreau's analytical model [5]. We use load distribution as a metric to evaluate the load balancing.

2) Average end-to-end delay: The end-to-end delay is the average time for all surviving data packets from the source to the destination.

3) The packet delivery fraction: The packet delivery fraction represents the percentage of the number of successfully received data packets at the destination to the number of data packets created by the source.

- Performance Evaluation in terms of Load Balancing

Fig. 3 portrays the load distribution of the two protocols as function of the distance from the network center. In our simulation, the center is the midpoint of the segment between source and sink nodes. With the increase of the distance to the center, there is a much more slight decrease of the load for the multipath routing while the load is greatly reduced for the shortest path routing. This simulation shows that our multipath routing can achieve better load balancing. This result can be explained by the fact that the traffic of the network is evenly regulated to the different paths while the single path always chooses the geographic-based shortest path, which will unfairly distribute more loads to the nodes along this optimal route than their neighboring nodes. According to this Figure, we conclude that the shortest path is likely to be overloaded because this route is across or very close to the center. In addition, due to the fact that we adopt load balancing policy, theoretically, all the nodes should experience approximately the same loads in the multipath routing 1, however, there exists a smooth decrease of the loads as the distance increases. The possible reason is that those packets that travel through longer routes are dropped due to more latency.

Still, we notice that as the distance from the network center increases, the number of average load for multipath routing 2 drops faster than multipath routing 1. This can be explained that multipath routing 2 is to choose a path with a probability inversely proportional to the length of the path. In other words, the further the distance to the center, the lower the probability that the nodes are used to relay the packets. Moreover, since our load balancing policy is not optimal, those nodes close to the optimal route have to be assigned more traffic in comparison with ones at the rear. Nevertheless, it is important to stress the fact that our multipath routing outperforms the shortest path routing in terms of load balancing.

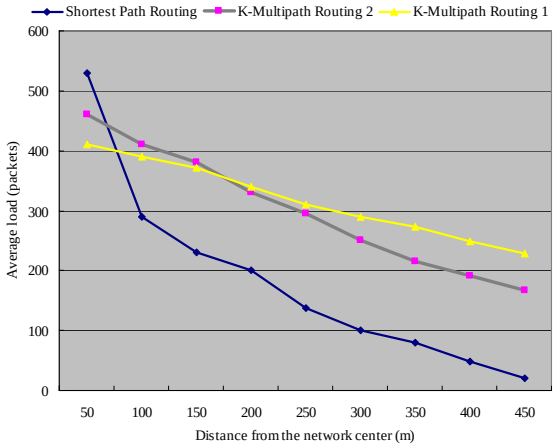


Figure 3. The average load distribution as function of distance from the network center.

- Performance Evaluation in terms of End-to-end Delay

Fig. 4 illustrates the average end-to-end delay as function of the sampling interval. As can be seen in Fig. 4, the end-to-end delay is greatly reduced for both k-multipath routing and the shortest path routing as the sampling interval increases, and compared with SPR, the advantage of k-multipath routing is obvious at high transmission rate. Multipath routing can reduce the queue delay because the total traffic load is distributed evenly into multiple routes, and therefore, data packets will experience less queuing delay. On the contrary, all the traffic would route along only one path, corresponding to the heavily congested path case. The benefits of multipath routing in reducing queuing delay is more prominent at smaller sampling interval networks where congestion has higher probability to occur. With the increase of the sampling interval, the traffic rate created with sampling interval becomes lower and the shortest path routing obtain better end-to-end delay performance than multipath routing. This can be explained by noting that nodes have sufficient time to process and distribute packets in timely manner in SPR at bigger sampling interval while more paths will be handled for multipath routing and thus the queuing delay of the data packets in the source node increases which leads to the increase of the average end-to-end delay. From the Fig. 4, the benefit of using multipath routing no longer holds when the sampling interval is more than 0.2 (<20 packets/s).

We also notice that k-multipath routing 2 performs better than k-multipath routing 1. This is because those longer length paths have longer delay, and in consequence, the routing protocol with a policy that less

traffic should be dispatched onto those longer length paths has lower mean delay for the whole network, compared to that policy that traffic is distributed more different paths fairly. Therefore, although through dispersing the load among multiple paths, we can improve end-to-end delay on heavily loaded networks especially on heavily loaded networks, we need to consider the impact of the strategy that how to allocate the traffic in order to achieve a minimum mean delay for the whole network.

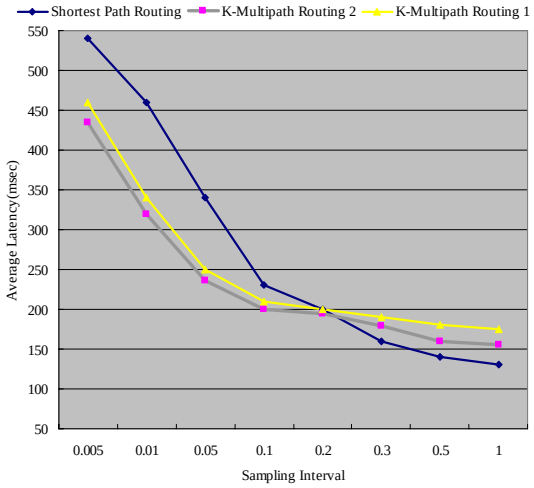


Figure 4. The average end-to-end delay as function of sampling interval.

- Performance Evaluation in terms of Packet Delivery Fraction

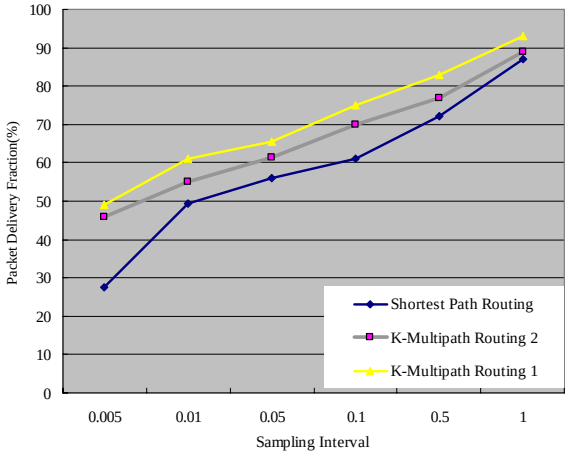


Figure 5. The packet delivery fraction as function of sampling interval.

The last experiment that we conduct is used to measure the packet delivery fraction of the network. Fig. 5 shows the packet delivery fraction comparison between SPR and the k-multipath routing. Clearly, k-multipath routing has a higher packet deliver fraction than SPR and with the increase of the sampling interval while the packet delivery fraction for all protocols appears to be increasing steadily. This is because as the sampling interval increases, the traffic rates inversely decrease, which in turn results in lower probability that they might be dropped due to congested areas. We still notice that the packet delivery fraction for the shortest path routing is smallest while the multipath routing 1 is largest. Since the loads are evenly distributed onto the multiple paths, there is less likely to be dropped for packets than the shortest path routing and the multipath routing that chooses a path based on its length due to the traffic congestion. As for the shortest path routing, it always uses the optimal route from a source to a destination. When the traffic rate increases to some threshold, the congestion becomes problematic.

VI. CONCLUSIONS

We present a novel load-balancing mechanism for wireless sensor networks. The new scheme is simple but very effective to achieve load balancing and congestion alleviation. We have explored an experimental comparison between k-multipath routing and the shortest path routing. Our performance study shows that

1) The network traffic can be distributed more evenly onto multipath routing. Load balancing is important to fairly distribute the routing task among the nodes of the network. It can also protect a node from failure considering that a node with heavy duty is likely to deplete its energy quickly.
2) The k-multipath routing can gain somewhat improvement of the average end-to-end delay at a higher traffic rate.
3) Although it takes much more time for the packet delivery along those multiple paths than the shortest path, the packet delivery fraction of our technique has been improved obviously and the network resource can be utilized efficiently.

ACKNOWLEDGMENT

This work has been supported by Scientific Research Fund of Computer School of Wuhan University under grant 2007AA01Z138. The authors would like to thank the reviewers for their detailed comments on earlier versions of this paper.

REFERENCES

[1] G. Oliviu Ghica, Goce Trajcevski, Peter Scheuermann, Zachary Bischoff, and Nikolay Valtchanov, SIDnet-SWANS: A simulator and integrated development platform for sensor networks application. Technical Report NWU-EECS-08-05, 2008.

[2] X.Chen, M.Chamania, A. Jukan, A. C. Drummond, N. L. S. da, Fonseca, QoS-Constrained Multi-path Routing for High-End Network Applications. IEEE INFOCOM2009 High-Speed Networks Workshop, Rio de Janeiro, Brazil , April 2009.

[3] J. Broch, D. Johnson, and D. Maltz, The Dynamic Source Routing Protocol for Mobile Ad hoc Networks, http://www.ietf.org/internet-drafs/draft-ierf-manet-dsr-04.txt, Nov. 2000, IETF Internet Draft.

[4] Perkins, C., Royer, E., Ad-hoc on-demand distance vector routing. in Proc. of the 2nd IEEE Workshop on Mobile Computing System and Applications, pp. 90–100 New Orleans, LA, USA, 1999.

[5] P.P. Pham and S. Perreau, Increasing the network performance using multipath routing mechanism with load balance, Ad Hoc Networks, vol.2, pp.433-459, 2004.

[6] M. Pearlman, Z. Haas, P. Sholander, and S. S. Tabrizi, On the Impact of Alternate Path Routing for Load Balancing in Mobile Ad-Hoc Networks, MobiHoc'2000, Boston, USA, August 11, 2000.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 389-392

Design of the Evolutional Group Buying Auction in Business to Business Electronic Commerce

Huafeng Li [1], and Yueting Chai[2]
[1] Department of Automation, Tsinghua University , Beijing, China
Email: li-hf07@mails.tinghua.edu.cn
[2] Department of Automation, Tsinghua University , Beijing, China
Email: chaiyt@ mail.tinghua.edu.cn

***Abstract*—Group buying auction(GBA) is seen as an effective form of B2C electronic commerce. GBA could not be directly applied to B2B transactions. This paper establish the evolutional group buying auction (EGBA) for B2B commodity trading based on GBA. First, the EGBA model is presented, and the key algorithms are introduced. Then Complicity Problem in EGBA is discussed . Finally, we validate the EGBA model through some examples.**

***Index Terms*—electronic commerce; business to business; online auction; evolutional group buying**

I. INTRODUCTION

The Internet's computational power and flexibility have made auctions a widespread and integral part of both consumer and business markets [1]. IN recent years, online auctions are popular both in the business-to-business (B2B) and the consumer markets. Auction played an important role in commerce as an effective price discovering mechanism. Johnson et. al. [2] point out that the online consumer auction sales in the US will reach $65 billion by 2010, accounting for nearly one-fifth of all online retail sales.

The traditional group buying auction (GBA) mainly aimed at B2C transactions [3, 4, 5]. Buyers can only bid for one commodity, and as the network characteristics, it may appear that many bidders successfully bid the same commodity at the last-minute, but GBA did not had good solution to this shortcoming. So, due to model deficiencies, GBA could not be directly applied to B2B transactions. Here, the evolutional group buying auction (EGBA) mainly aims at B2B commodity trading, which is the improvement based on group buying auction (GBA).

The rest of the paper is organized as follows. In Section 2, the EGBA model is presented, and the key Algorithms are introduced. Complicity Problem in EGBA is discussed in Section 3. We validate the EGBA model and give some examples in Section 4. Finally, Section 6 draws some conclusions and future work.

II. EGBA MECHANISM DESIGN

A. Model description

EGBA is an extension of the traditional discount sales methods, which was based on homogeneous multi-items auction. The popularity of the traditional quantity discount has been studied thoroughly [6, 7, 8]. All bidders were a group. The more bidders, the more numbers of goods were sold at the lower the price , so EGBA suitable for large B2B commodity trading. The basic EGBA model is established as the following:

There are one seller and n-buyers. Sellers set the price ladder in accordance with the quantity commodity quantity, and each buyer is an independent bid. The seller is auctioning a total of L items. The auction start time is recorded as 0, and set the auction end time which is recorded as T . The auction is ended according to Rule 2 and Rule 3 in the following parts.

Set on the amount purchased by the seller in accordance with the price ladder, ladder price vector which the seller sets is denoted by

$$S = \begin{pmatrix} s_1 & s_2 & \dots & s_i & \dots & s_m \\ l_1 & l_2 & \dots & l_i & \dots & l_m \end{pmatrix}^{\mathrm{T}}$$

Where $i = 1, 2, ..., m$. The price vector satisfied $s_1 > s_2 > ... > s_m$ and $0 < l_1 < l_2 < ... < l_m \le L$. Row vector (s_i, l_i) indicated that the total purchase amount Y to satisfy $l_i \le Y < l_{i+1}$ at the price of s_i .

Each buyer at moment t_j bid b_j and decided to purchase the number of x_j . All buyer's bid vector is denoted by

$$B = \begin{pmatrix} t_1 & t_2 & \dots & t_j & \dots & t_n \\ b_1 & b_2 & \dots & b_j & \dots & b_n \\ x_1 & x_2 & \dots & x_j & \dots & x_n \end{pmatrix}^{\mathrm{T}}$$

Where $j = 1, 2, ..., n$, that a total of n of buyers bid, and $t_j \in (0, T)$. The total amount of goods are recorded as $X = \sum_{j=1}^{n} x_j$ (Not necessarily equal to the last traded quantity).

The EGBA sets the relevant rules as follows:

Rule 1 Priority principle. In the following three cases, the buyer i gives priority to the seller j access to goods: (1) $t_i < t_j$;(2) $t_i = t_j, x_i > x_j$;(3) $t_i = t_j, \quad x_i = x_j,$

$b_i > b_j$. That is in accordance with the time priority, the number of priority, and the price priority.

Rule 2 The first condition to end the auction. Under the conditions of the rule 1, the auction time is taken to reach T, then the auction is ended.

Rule 3 The second t condition to end the auction. In time T, under the conditions of the rule 1, if the $\mathrm{X} \geq L$ then the auction is also ended at this time.

B. Algorithm Design

In the EGBA auction mechanism, in the number of bidders to determine the circumstances, the need to determine the final transaction price and transaction volume, as well as the winning buyer. Here the algorithm which we design is to determine the final transaction price, the successful bidders set and the winning number of commodities.

1) Determine the transaction price algorithm

Denote the ransaction price based on B for $\mathrm{P}(B)$ .Then, $\mathrm{P}(B) \in \{s_1, s_2, ..., s_i, ..., s_m\}$. In order to determine the final transaction price, first of all defined functions $\Phi(\cdot)$:

$$\Phi(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

Consider

$$\begin{aligned} & i^* = \arg(\max \sum_{j=1}^{n} x_j \cdot \Phi(b_j - s_i)) \\ & s.t \ \sum_{j=1}^{n} x_j \cdot \Phi(b_j - s_i) \geq l_i \ , \quad \forall i \in [1, m] \end{aligned} \tag{1}$$

Denote the target subscript for i^* , then the final transaction price $\mathrm{P}(B) = s_{i^*}$, corresponding to the price ladder (s_{i^*}, l_{i^*}) .

Specially, in accordance with Rule 2 of the end of the auction, the transaction volume is L . It is satisfied $L \geq l_m$, so that $\mathrm{P}(B) = s_m$. By the above algorithm (1) we can also get the same results.

2) The algorithm to calculate the successful bidder set and the corresponding winning number of commodities.

In the circumstances of getting the transaction price $\mathrm{P}(B)$, the next step is to determine the successful bidder set and the corresponding winning number of commodities. In order to facilitate the expression, we denote the corresponding successful bidder's subscript set for J .

According to the different auction ending conditions, it can be divided into two kinds of calculation:

a) In accordance with Rule 2 to end the auction

In this case, the auction time reached T . It is clearly $\mathrm{X} < L$ Therefore the tender price which was greater than the transaction price for each buyer can receive the desired number of commodities. The tenderer's set J could be obtained using the following formula:

$$J = \{ j \mid b_j - \mathrm{P}(B) \geq 0, \ j \in [1, n] \} \tag{2}$$

The final successful bidder j got x_j units commodities , Where $j \in J$.

b) In accordance with Rule 3 to end the auction

In this case, the tender commodities was greater than or equal to the number of the total number of auction items, namely $\mathrm{X} \geq L$. Here comes down to how to allocate the commodities. Then we must follow the rules 1 to re-row the row vector of the bid vector B , and one by one according to the new order to allocate goods. That is, at first in accordance with the tender t_j re-arranged from small to large order, then if the times was the same , we re-arranged in accordance with x_j descending order. If both the time and the purchase of the same amount were the same, we re-arranged by b_j with descending order. Finally, we get the re-arranged bid vector

$$B' = \begin{pmatrix} t_{k_1} & t_{k_2} & \cdots & t_{k_j} & \cdots & t_{k_n} \\ b_{k_1} & b_{k_2} & \cdots & b_{k_j} & \cdots & b_{k_n} \\ x_{k_1} & x_{k_2} & \cdots & x_{k_j} & \cdots & x_{k_n} \end{pmatrix}^{\mathrm{T}} \tag{3}$$

According to the row sequence of vector B' , the successful bidder from the first line start in this order, for example, k_1 is the first successful, and the second one is the k_2 , and so on. Denote the last row of the successful bidder by r^* , then last one successful bidder is k_{r^*} . It can be obtained using the following formula:

$$\begin{aligned} & r^* = \arg(\min \sum_{j=1}^{r} x_{k_j}) \\ & s.t \ \sum_{j=1}^{r} x_{k_j} \geq L \ , \quad \forall r \in [1, n] \end{aligned} \tag{4}$$

A collection of the successful bidder for the

$$J = \{ k_j \mid j \in [1, r^*] \} \tag{5}$$

In this case, not all of the successful tenderer can obtain the desired number of goods, specifically, the last one successful bidder k_{r^*} may not get the amount of the bid amount of goods. Denote the last one successful bidder quantity of goods the last one successful bidder got by x_{r^*}' , then $x_{r^*}' \leq x_{r^*}$. We can get

$$x_{r^*}' = L - \sum_{j=1}^{r^*-1} x_{k_j} \tag{6}$$

Aside from the successful bidder k_{r^*} , The remaining quantity of goods was distributed to the other successful bidders for the x_{k_j} , where

$$k_j \in J' = \{k_j \mid j \in [1, r^* - 1]\}.$$

c) Algorithmic process

According to auction rules and the above algorithm we can get the specific algorithm processes in the following:

Step 1 To determine whether the auction time to reach T, if the time taken to reach T, the auction ended, and if not, skip step 2 to step 3;

Step 2 Calculate $X = \sum_{j=1}^{n} x_j$. To determine whether $X \geq L$? If met, then the auctions have ended, go to step 5. If the auction does not end, then jump to Step 1;

Step 3 Calculate $P(B)$ according to the formula (1);

Step 4 Calculate J according to the formula (2) if the auction was ended at Rule 2. All the buyers in set J are successful bidders, the winning number of commodities for their own bid amount x_j ($j \in J$). Go to step 7;

Step 5 If the auction was ended at Rule 3 , then $P(B) = s_m$. According to the formula (3) get the re-arranged bid vector B'.

Step 6 After getting B', according to the formula (4) to calculate the last one successful bidder in the line r^* , then the collection of the successful bidder $J = \{k_j \mid j \in [1, r^*]\}$. According to the formula calculation (6) to obtain x_{r^*}', the remaining amount of the winning buyers got bids when the number of commodities were x_{k_j}, where

$$k_j \in J' = \{k_j \mid j \in [1, r^* - 1]\};$$

Step 7 end.

III. COMPLICITY PROBLEM

Conspiracy is a common phenomenon in the auction, and conspiracy, some or all of the co-ordination among the bidders bid their own actions in order to get in the auction, when acting alone higher than the respective gains. Conspiracy may appear to be a clear agreement: Which field in which bidders win the auction, it may appear to be consistent bidders were secretly down their bids. In order to facilitate the description of the bidders valuation of goods, as well as the relationship with bid price and the price ladder, we define the function $\theta(\cdot)$:

$$\theta(v) = \begin{cases} s_1 & v \geq s_1 \\ s_i & s_i \leq v < s_{i-1} \\ s_m & v = s_m \\ 0 & v < s_m \end{cases}$$

EGBA complicity in the auction will be met by Theorem 1 and Theorem 2 to describe.

Theorem 1 In EGBA mechanism, the conspiracy within the group of bidders will not reduce the tender price.

Proof:

q bidders form a conspiracy group ($1 < q \leq n$, n is the total number of bidders), where the conspirators set denoted $\Omega_q = \{k_j \mid j \in [1, q]\}$. Member of the complicity in the bidding vector B Where in the line k_j . The Group's bid collusion with the total number of goods Expressed as X_{Ω_r} , $X_{\Omega_q} = \sum_{j \in \Omega_q} x_j$, The seller set the corresponding price vector for the (s_i, l_i) . That satisfy $l_i \leq X_{\Omega_r} < l_{i+1}$. And seek a final transaction price under $P(B) \leq s_i$. Because they do not complicity of the tender price $b_j = \theta(v_j) \geq s_i$ (Because if $\theta(v_j) < s_i$, The $j \in \Omega_r$ number of other bidders in uncertain circumstances, the conspirators of the final winning bid gains may be negative) otherwise they would not form a conspiracy group, according to auction rules in the price of a principle of giving priority conspirator in determining the $P(B) \leq s_i$. Will not reduce the tender price, otherwise you will reduce the possibility of winning. Q.E.D.

Theorem 2 In EGBA mechanism, the buyer's complicity will not damage the interests of the seller.

Proof:

According to Theorem 1, the conspirators will not reduce the tender. Therefore, according to formula (1) the final transaction price will not be reduced. And then according to formula (2) ~(6) we can see the number of goods in a conspiracy case the ultimate buyer will not reduce the total number of successful products (or even likely to increase). So the buyer's complicity will not damage the interests of the seller. Q.E.D.

IV. VALIDATION OF THE MODEL

Suppose that one seller has a total of $L = 300$ tons commodities for sale, and the auction time is set to $T = 5$. Ladder price vector is

$$S = \begin{pmatrix} 1000 & 900 & 800 & 700 \\ 1 & 51 & 101 & 201 \end{pmatrix}^T$$

It can be described as Figure 1.

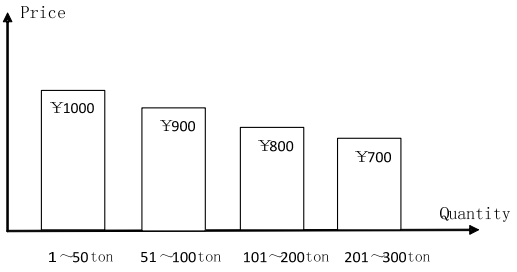


Figure 1. Examples

A. The example to end the auction in accordance with Rule 2

At the end of time corresponding to all buyers bidding vector is as follows:

$$B=\begin{pmatrix} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \\ b_1 & b_2 & b_3 & b_4 & b_5 & b_6 \\ x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{pmatrix}^{\mathrm{T}}$$

$$=\begin{pmatrix} 1 & 1.5 & 2 & 2.5 & 3 & 4 \\ 1000 & 900 & 900 & 800 & 700 & 700 \\ 40 & 60 & 10 & 20 & 20 & 10 \end{pmatrix}^{\mathrm{T}}$$

Since the end of the auction in accordance with Rule 2, directly in accordance with the algorithmic process steps 3 and 4 can be calculated,

Step 3: Calculate the transaction price, according to equation (1):

$$i^*=\arg(\max\sum_{j=1}^{n}x_j\cdot\Phi(b_j-s_i))$$

$$s.t\ \sum_{j=1}^{n}x_j\cdot\Phi(b_j-s_i)\ge l_i\ ,\quad \forall i\in[1,m]$$

We can get $i^*=3$, and $\mathrm{P}(B)=s_3=800$.

Step 4: Calculate the successful bidder collection ,

$$J=\{j\mid b_j-80\ge 0,\ j\in[1,6]\}=\{1,2,3,4\}\ ,$$

That is, 1 to 4 winning bidder to obtain the number of goods followed by 40, 60, 10, 20 .

B. The example to end the auction in accordance with Rule 3

Reach the time $t_7=3.5$. Vector corresponding to the tender is as follows:

$$B=\begin{pmatrix} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 \\ b_1 & b_2 & b_3 & b_4 & b_5 & b_6 & b_7 \\ x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \end{pmatrix}^{\mathrm{T}}$$

$$=\begin{pmatrix} 1 & 1.5 & 2.5 & 2.5 & 3.5 & 3.5 & 3.5 \\ 700 & 900 & 800 & 700 & 700 & 800 & 800 \\ 40 & 80 & 50 & 80 & 40 & 40 & 20 \end{pmatrix}^{\mathrm{T}}$$

According to algorithmic process in step 2,

$$\mathrm{X}=\sum_{j=1}^{6}x_j=320>L\ ,$$

The auction ended, and then follow steps 5 and 6 :

Step 5: We got $\mathrm{P}(B)=s_m=700$, and the Re-arranged bid vector was

$$B'=\begin{pmatrix} t_1 & t_2 & t_4 & t_3 & t_6 & t_5 & t_7 \\ b_1 & b_2 & b_4 & b_3 & b_6 & b_5 & b_7 \\ x_1 & x_2 & x_4 & x_3 & x_6 & x_5 & x_7 \end{pmatrix}^{\mathrm{T}}$$

$$=\begin{pmatrix} 1 & 1.5 & 2.5 & 2.5 & 3.5 & 3.5 & 3.5 \\ 700 & 900 & 700 & 800 & 800 & 700 & 800 \\ 40 & 80 & 80 & 50 & 40 & 40 & 20 \end{pmatrix}^{\mathrm{T}}$$

Step 6: According to the formula (4), we got $r^*=6$ and $J=\{k_j\mid j\in[1,r^*]\}=\{1,2,4,3,6,5\}$. So successful bidder 1, 2, 4, 3, 6 to obtain the number of goods 40, 80, 80, 50, 40, and the last one successful bidder $k_{r^*}=5$ got the number

$$x_{r^*}{}'=L-\sum_{j=1}^{r^*-1}x_{k_j}=300-290=10$$

V. CONCLUSIONS

In this paper, we design the evolutional group buying auction for business to business electronic commerce, which is the improvement of the traditional evolutional group buying auction. The basic model is firstly established. Then we design several algorithms to determine the final transaction price, the successful bidders set and the winning number of commodities. At last, some examples are given to validate the above model.

REFERENCES

[1] Edieal J. Pinker, Abraham Seidmann, and Yaniv Vakrat, "Managing Online Auctions: Current Business and Research Issues," Management Science, vol.49, No.11, ,Movember 2003, pp.1457-1484

[2] A.C. Johnson, B. Tesch, "A Forecast and analysis of US auction Sales to consumers," Forrester Research, 2005.

[3] Jian Chen, Xilong Chen, and Xiping Song, "Bidder's Strategy Under Group-Buying Auction on the Internet",IEEE Transactions on Systems, Man and Cybernetics—Part A: Systems and Humans, vol. 32, NO. 6, November 2002

[4] Jian Chen, Xilong Chen, and Xiping Song,"Comparison of the group-buying auction and the fixed pricing mechanism",Decision Support Systems,2007 , pp. 445–459

[5] Jian Chen, Xilong Chen, and Robert J. Kauffman, "Cooperation in group-buying auctions", Proceedings of the 39th Hawaii International Conference on System Sciences,2006

[6] Kohli, R., and P. Heungsoo. 1989. "A cooperative game theory model of quantity discounts", Mgmt. Sci.,35(6),pp. 693-707.

[7] Lee, H.L. and R. Meir, 1986. "A generalized quantity discount pricing model to increase supplier's profits,"Mgmt. Sci., 32(9), pp.1177-1185.

[8] Monahan, J.P. 1984. "A quantity discount pricing model to increase vendor profits," Mgmt. Sci. 30(6),pp.720-726.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 393-397

Motion Sequence Filtering using Geometric Algebra

Zhixin Xue , Shi Wan , and Jiawen Zhou
Nanchang University/ Dept. of Computer Science, Nanchang, China
Email:xue2zhixin@yahoo.com.cn

***Abstract*—This paper presents novel methods of filtering motion sequence, which extend image processing algorithms to deal with motion sequences. The extending approach is the first to processing motion sequences represented by geometric algebra. It maintains many desired properties of geometric algebra such as compactness, generality and intuitive meaning. Experiments with captured motion sequences showed that presented methods are effective and can be useful in many aspects of motion processing.**

***Index Terms*—geometric algebra, image processing, heat diffusion equation**

I. Introduction

Human motion data play a vital role in many research fields, like machine vision, robotics and computer animation. Many software and hardware packages have been designed to deal with human motion capturing. However, capturing subtle details and different kinds of human motion is still a challenging task for many researchers. The difficulty comes from two factors: data obtained during capturing are highly noisy due to hardware deficiencies, and processing such multidimensional data is different from standard signal processing work. Motion data captured in real environment have built-in space-time dependency. Improper signal processing can bring about violation of both human structure and physical laws. The solution could be either to take into consideration all forces and real environment constraints or to use approaches which do not bring in skeleton structure modification. The first method is difficult to achieve in that to create a realistic animation covering all actual constraints in an analytical way is nearly impossible. In this paper we will show that the second method is able to be partially achieved by using correct data representation and selecting suitable de-noising algorithms.

Motion sequence filtering has been intensively researched from the late 1980s. J. Lee and S. Y. Shin in Ref. [5] presented a motion fairing method to deal with both rotational and translational motions. In Ref. [6], energy criterion and minimization schemes were constructed in order to obtain a filtered motion sequences. J. Lee and S. Y. Shin in Ref. [5] successfully applied this idea in motion blending. The research in Ref. [6] proposed concepts to construct smooth interpolation for filtering motion sequences. Different kinds of equations (parabolic, hyperbolic and elliptic) can be used for motion sequence filtering. The method was extended in Ref. [9], where a nonlinear coefficient into the equation was introduced. It was afterwards studied in detail in Ref. [10] and also generalized as a geometrical framework in Ref. [10]. In cited papers, some general approaches for filtering motion sequence were introduced. Our work focuses on adopting some of those schemes for motion data represented by geometric algebra.

The rest of the paper is organized as follows: Section II gives a brief introduction to geometric diffusion model for image processing. Our geometric algebra approach is given in Section Ⅲ. The geometric diffusion methods for motion data are provided in Section IV. Experimental results are presented in Section V, following conclusions and future work.

II. Geometric Diffusion Model For Image Processing

In this section, heat diffusion equation is introduced to remove noise from an image by modifying the image with the least undesired deformations. Consider the isotropic heat diffusion equation.

$$\frac{\partial I(x,y,t)}{\partial t} = div(\nabla I) \tag{1}$$

with the original image *I(x,y,0)* as the initial condition, where $I(x,y,0):R^2 \rightarrow R$ is an image in the continuous domain, *(x,y)* specify spatial position, *t* is an artificial time parameter. The above parabolic type equation is utilized in Ref. [2] for image restoration. Many researchers have confirmed that parabolic equations give satisfying results in image processing. Running the heat diffusion process with an initial value function, we can get the whole family of images for consecutive moments. As it was shown in Ref. [9], we treat images as scale-space functions, which are solutions of initial-value equation in time *t*. One of the most popular ways is to filter with a radial Gaussian kernel, which filtering is called Gaussian blur. In fact, Gaussian blur is computed in discrete domain, which offers a good approximation of the continuous domain. Although Gaussian blur is very good for local reducing noise, this filter also destroys the image content. Perona and Malik replaced the classical isotropic diffusion equation with the following:

$$\frac{\partial I(x,y,t)}{\partial t} = div[g(\|\nabla I\|)\nabla I] \tag{2}$$

where $\|\nabla I\|$ is the gradient magnitude and $g(\|\nabla I\|)$ is an edge-stopping function. Although many types of stopping functions were extensively researched, only two

AP-PROC-CS-09CN005

types of edge-stopping function *g* are considered in our project, which satisfy $g(x) \to 0$ when $x \to \infty$ so that the diffusion is stopped across edges:

$$g(x) = e^{-(x/K)^2} \tag{3}$$

$$g(x) = \frac{1}{1+(x/K)^2} \tag{4}$$

where K is an extra parameter which controls the heat diffusion process. This anisotropic diffusion process is highly nonlinear. And it is impossible to get an analytical solution for anisotropic diffusion process. Its approximate solution can be obtained by discretizing between 4-nearst neighbors of domain:

$$I_{i,j}^{(t+1)} = I_{i,j}^{(t)} + \lambda(c_N d_N + + c_E d_E + c_S d_S + c_W d_W) \tag{5}$$

where $I_{i,j}^{(t)}$ is discretely sampled image, subscripts *(i,j)* denote the pixel position in a discrete 2D grid, and t denotes discrete time steps. c_η and d_η (η is the direction of four neighbors.) simply define local differences in four directions. Parameter λ is a scalar that determines the rate of heat diffusion. The last *c* and *d* variables are defined in the following way:

$$d_N = I_{i-1,j} - I_{i,j}, c_N = g((\nabla I)^t_{i-1/2,j}) \approx g(d_N) \tag{6}$$

$$d_E = I_{i,j+1} - I_{i,j}, c_E = g((\nabla I)^t_{i,j+1/2}) \approx g(d_E) \tag{7}$$

$$d_W = I_{i,j-1} - I_{i,j}, c_W = g((\nabla I)^t_{i,j-1/2}) \approx g(d_W) \tag{8}$$

$$d_S = I_{i+1,j} - I_{i,j}, c_S = g((\nabla I)^t_{i+1/2,j}) \approx g(d_S) \tag{9}$$

Experiments and theoretical analysis confirmed that this method offers good results for image filtering. And discretization procedure does not produce serious instability effects, so it encourages us to select this approach for motion data filtering.

III. Motion Model Of Geometric Algebra

The model of a human body in our project is composed of 23 bones arranged in a rigid hierarchy. Each human body has its virtual root bone, which is the origin of the whole hierarchy. The actual position of each bone relies upon positions of all its predecessors in the kinematical chain. In practice, only the rotation data of each bone are recorded. Moreover, the data about global translation of the human body is specified, it is generally only for the virtual root bone. In computer animation, there are three representation methods for rotations: Euler's angles, quaternions and geometric algebra. Geometric algebra provides compact representation and works well in any other dimension without changes.

Geometric algebra (Clifford algebra) is a powerful new computing paradigm for many research fields such as computer graphics, CAD/CAM, robotics and machine vision. It offers a coordinate-free approach to model geometric objects and simple symbolic algorithm. As a unifying computing method, it comprises a number of geometric descriptions widely used in computer graphics and mechanical engineering. A short introduction of GA (Geometric algebra) is given in the next. More detailed materials can be found in Ref. [1] and Ref. [4].

GA (Geometric algebra) generalizes linear algebra concepts by introducing two-, three-, or higher-dimensional subspaces, called blades. In GA, vectors can be combined using the outer product to obtain higher-dimensional entities, such as bivectors (representing planes) and trivectors (representing 3D subspaces). Given two vectors *a* and *b*, their outer product $a \wedge b$ is a blade of grade 2 (or a 2-blade), which represents the two-dimensional oriented subspace that contains *a* and *b*. Such blades of grade 2 are called bivectors. The outer product of three vectors results in a blade of grade 3 (also called a trivector), and so on. According to this formalism, scalars are 0-blades, and vectors are 1-blades. GA also defines the operators with which these subspaces can be manipulated. It is possible to add and subtract subspaces of different dimensions by means of composition, and even to multiply them, resulting in powerful expressions that can describe many geometric relations and concepts.

Considering an orthogonal basis $\{e_1, e_2, e_3\}$ of Euclidian space R^3, a possible basis of the 3-dimensional GA, Cl_3 consists of all 0-dimensional subspaces (scalars, 0-blades), 1-dimensional subspaces (vectors, 1-blades), 2-dimensional subspaces (bivectors, 2-blades) and 3-dimensional subspaces (trivectors, 3-blades), which are listed in Table Ⅰ. The total number of basic k-blades in these subspaces is 8. The blade of the highest dimension is called the pseudoscalar. A multivector in Cl_3 is a linear combination of different k-blades. In R^3, it contains a scalar part, a vector part, a bivector part and trivector part, which can be defined with 8 real numbers: $\alpha_1 + \alpha_2 e_1 + \alpha_3 e_2 + \alpha_4 e_3 + \alpha_5 e_1 e_2 + \alpha_6 e_2 e_3 + \alpha_7 e_1 e_3 + \alpha_8 e_1 e_2 e_3$.

The most important operator of GA is the geometric product between two multivectors, which combines the outer product with the familiar dot product. For two vectors *a* and *b*, the geometric product can be calculated: $ab = a \bullet b + a \wedge b$, and the resulting terms can be simplified by means of the set of axioms listed in Table Ⅱ.

A special type of multivector is the spinor, a scalar element adding a bivector element, $R = \cos\frac{1}{2}\theta + \sin\frac{1}{2}\theta A$. The spinor can perform rotations of vector *v* to the *A* plane, by doing $v' = R^+ v R$, where R^+ is the reverse of R. In other words, the spinor performs a rotation around the normal of the A plane, denoted by the A^+.

The full description of the motion sequences is given by the matrix of dimensions $k \times n$. We postulate that the k is fixed to 23 which is the number of skeleton bones. The length of the animation is *n*, and it relies on the actual sequence. Each element in the matrix records a spinor R, which represents rotation between the initial position of the bone k and the position in frame t.

TABLE I.
BLADES IN 3-DIMENSIONAL GEOMETRIC ALGEBRA

Dimension	Element	Blade
0	Scalar	1
1	Vector	e_1, e_2, e_3
2	Bivector	e_1e_2, e_2e_3, e_1e_3
3	Trivector	$e_1e_2e_3$

TABLE II .
GEOMETRIC PRODUCT AXIONS IN CL_3

$e_i\, e_i = 1$	i=1,2,3
$e_i\, e_{j=-}\, e_j\, e_i$	$i \neq j$=1,2,3
$\lambda\, e_i = e_i\, \lambda$	i=1,2,3

IV. GEOMETRIC DIFFUSION MODEL FOR MOTION DATA

In implementing the filtering algorithms of motion data, the complexity of motion data often leads to violation of human structure. We suggest approaches of filtering which are derived from image processing algorithms, but they are revised for specific properties of motion data.

A. Gaussian Blur for GA Motion Data

As we saw in Section Ⅱ, the solution of isotropic diffusion equation is equivalent to the convolution with a Gaussian kernel. In practice, it is approximated by substituting the convolution with properly constructed interpolation of GA data. Considering a time-series of geometric algebra data *R[i]*, *i*=1,…, n, we yield two auxiliary geometric algebra data about sample *i*-th as follows:

$$R_p = R^{(t)}[i](R^{(t)}[i])^{-1} R^{(t)}[i-1])^{w} \tag{10}$$

$$R_n = R^{(t)}[i](R^{(t)}[i])^{-1} R^{(t)}[i+1])^{w} \tag{11}$$

where Rp is a geometric algebra data interpolated between considered geometric algebra data *R[i]* and the previous one in the sequence, and Rn denotes the result of the interpolation with the successor. Parameter w is interpreted as weight, it determines the strength of the interpolation. The actual result of processing the geometric algebra i-th frame data is calculated like:

$$R^{(t+1)}[i] = R_p((R_p)^{-1} R_n)^{1/2} \tag{12}$$

In order to acquire a filtered motion sequence, one has to proceed with the above operations for each frame from 2 to n-1.

B. Anisotropic Diffusion Equation for GA Motion Data

Experiments showed that Gaussian blur algorithms produce too smooth sequences, and some important features of the motion are lost in filtering noise. The opinion is to replace isotropic diffusion equation with anisotropic ones in the space of geometric algebra, which preserve edges while removing noise.

The 'edge' in the geometric algebra sequence is defined as a rapid change of neighboring positions, including directed or no directed change. We raise the interpolation scheme on the basis of expressions (6)-(9). The first step likes Gaussian blur approach:

$$r_P = R^{(t)}[i](R^{(t)+}[i]R^{(t)}[i-1])^{w_P(R)} \tag{13}$$

$$r_N = R^{(t)}[i](R^{(t)+}[i]R^{(t)}[i+1])^{w_N(R)} \tag{14}$$

where $R^{(t)+}$ denotes the reverse of $R^{(t)}$. The difference is that the parameter w is no longer constant. We will calculate two parameters by corresponding edge-stopping functions defined in (3) or (4).

$$w_P = g(|\, d(R[i], R[i-1])\,|) \tag{15}$$

$$w_N = g(|d(R[i], R[i+1]|) \tag{16}$$

The value of w_P and w_N is computed as follows:

$$d(R[i], R[i-1]) = \arccos(\mathrm{Re}((R[i])^{+} R[i-1])) \tag{17}$$

$$d(R[i], R[i+1]) = \arccos(\mathrm{Re}((R[i])^{+} R[i-1])) \tag{18}$$

where $(R[i])^{+}$ is the reverse of *(R[i])* and $\mathrm{Re}((R[i])^{+} R[i\text{-}1])$ denotes the scalar part of $((R[i])^{+} R[i\text{-}1])$. The last step is similar to Gaussian blur algorithm. We find a mean position between interpolated spinors and denote it as r_D for further computations:

$$r_D = r_P(r_P^{+} r_N)^{1/2} \tag{19}$$

In order to deal with the change of the motion direction, we suggest adding an extra step which takes into account direction of the filtered sequence. This approach is easy to compute and does not require extra assumptions on edge-stopping functions.

In the final step of the filtering, we define a new measurement *dir*(R[i]) called directional function, which should measure the local change of direction of the motion. In our project, we designed the following steps for directional function evaluation:

1) Compute a mean position for m predecessors and m successors from the given i-th frame.

$$r_{MP} = R[i-3]((R[i-3])^{+}(R[i-2]((R[i-2])^{+} R[i-1])^{1/2}))^{1/3} \tag{20}$$

The result are two spinors, respectively, r_{MP}, r_{MN}.

2) Compute the differences of direction between the given position ith and means just computed.

$$q_{dP} = (R[i])^{+} q_{MP} \tag{21}$$

$$q_{dN} = (R[i])^{+} q_{MN} \tag{22}$$

3) The value of directional function for position ith is defined as the distance between q_{dP} and q_{dN}.

$$dir(R[i]) = \arccos(\mathrm{Re}(q_{dP}^{+} q_{dN})) \tag{23}$$

4) In the end, we can implement the last step of the filtering. Using the result of r_D, we compute a new *R[i]* for the next time-step:

$$R^{(t+1)}[i] = R^{(t)}((R^{(t)}[i])^{+} r_D)^{w_D(R)} \quad (24)$$

To achieve a proper filtering, we can compute a variable weight $w_D(R)$ using the corresponding edge-stopping functions g like in (3) or (4):

$$w_D(R) = g(|\, dir(R[i])\,|) \quad (25)$$

The procedure should consider all motion data of given time-series. Boundary conditions like that mentioned in the Gaussian filtering. Note that due to the nonlinear character of the process, its properties concerning boundary elements will also change. Our algorithms are derived from anisotropic diffusion equation concepts for image filtering, but are revised for geometric algebra data.

V. Experiments And Discussion

Experiments were carried out to evaluate the performance of the presented methods. In the first group of experiments, walking motion sequences were artificially generated to test real performance of filtering algorithms. The second group is executed with some real captured running motion sequences. For comparison purposes, a randomly generated noise was contained in geometric algebra sequences. The filtering methods were implemented with the following parameters:

Gaussian blur filtering, with fixed weight coefficient $w=0.22$.

Anisotropic diffusion equation filtering, *g* function type 1, $\lambda_{dir}=0.23$, $\lambda=0.21$, $K=23$, $K_{dir}=0.63$.

From the test results, we observed that both methods can be applied to motion data filtering. Gaussian blur filtering provides faster filtering, however it brings about deformed sequences after a few iterations. On the contrary, anisotropic diffusion equation filtering produces better filtering results. What is more, motion sequences with rapid changes of motion direction are also preserved during the full iteration. The figures below give results of the algorithms for real motion data. The tested sequences were deformed with noise. The following algorithms were executed for sequence filtering:

Gaussian blur filtering, with fixed weight coefficient $w=0.11$.

Anisotropic diffusion equation filtering, g function type 2, $\lambda_{dir}=0.06$, $\lambda=0.16$, $K=24$, $K_{dir}=3$.

We noticed that g function of the second type provides better results for filtering real motion data. Results of two algorithms show that the quality and efficiency can be greatly improved by carefully choosing parameters. The parameter choice of the best results demands further researches. Our experiments found that anisotropic diffusion equation filtering presents better performances in the sequence much more deformed. Furthermore, we noticed that Gaussian blur gives too smooth sequence as compared with the original one. The results of the anisotropic diffusion equation filtering resemble the original sequence (compare locations of legs on all 6 figures). The features of human motion are preserved well after many iteration steps.

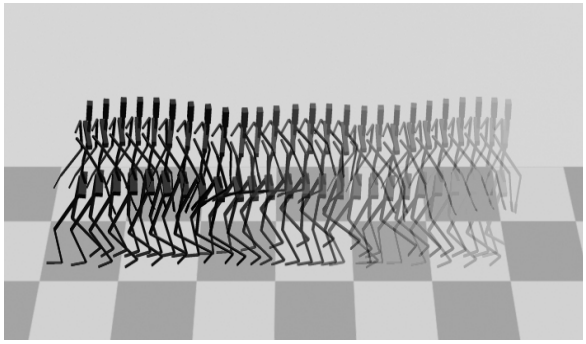


Figure.1 Walking motion sequences used to test the algorithms

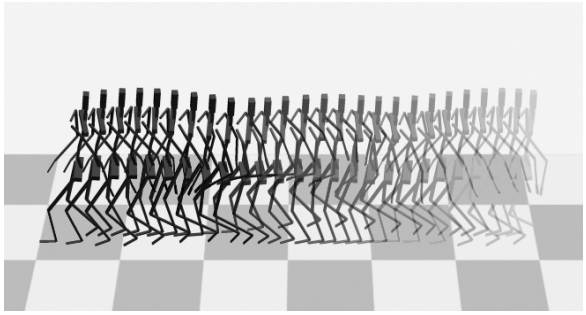


Figure.2 Anisotropic diffusion equation filtering result

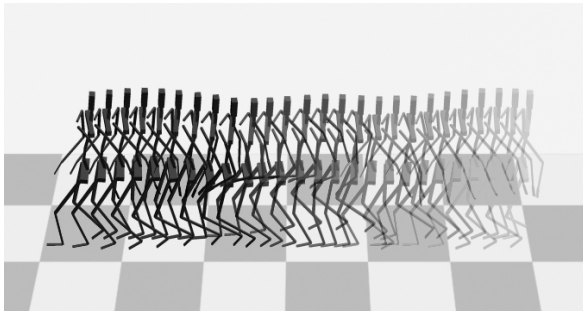


Figure.3 Gaussian blur filtering results.

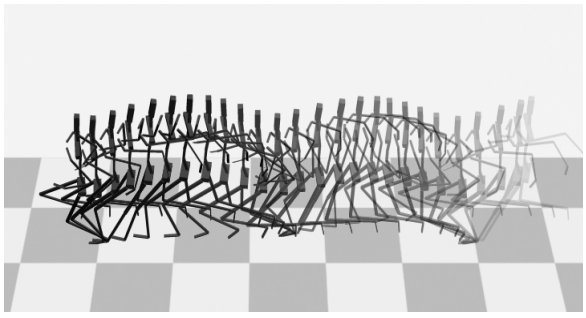


Figure.4 Running motion sequences used to test the algorithms.

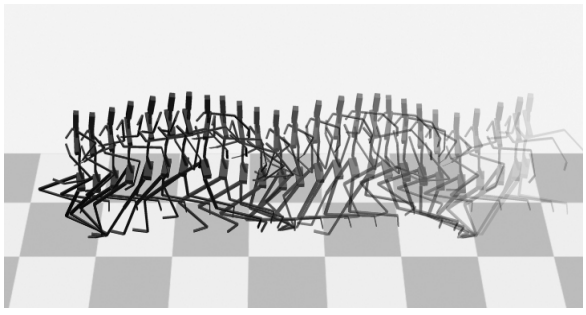


Figure.5 Anisotropic diffusion equation filtering result.

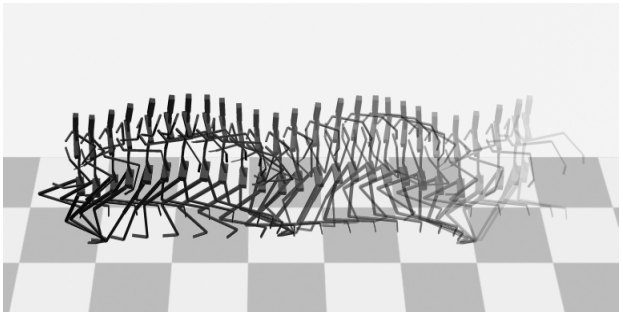


Figure.6 Gaussian blur filtering result.

VI. Conclusions And Future Work

Geometric algebra is a unified powerful analysis tool that offers an integrated approach to model kinetic phenomena. In this paper, we have presented two filtering methods using geometric algebra, which are extended from image processing algorithm of heat diffusion process. The first approach is very similar to methods proposed in Ref. [10]. The second approach, as far as we know, has not been previously presented with the application to geometric algebra data. Two methods are tested in experiments. The results are encouraging and provide a guidance mechanism to explore miscellaneous aspects of human motion processing.

References

[1] D. Hestenes, G. Sobcyk, *Clifford Algebra to Geometric Calculus: A Unified Language for Mathematics and Physics*, Kluwer Academic Publishers, 1987.

[2] Guillermo Sapiro, *Geometric Partial Differential Equations and Image Analysis*, Cambridge University Press The Edinburgh Building, Cambridge CB2 2RU, UK, 2001.

[3] Fletcher Dunn and Ian Parberry, *3D math primer for graphics and game development*, Jones & Bartlett Publishers, 2002.

[4] John Vince, *Geometric Algebra for Computer Graphics*, Springer-Verlag London Limited 2008.

[5] J. Lee, S. Y. Shin, "Motion fairing," *Computer Animation 96 Geneva, Swiss*, 1996, pp. 136–143.

[6] J. Lee, S.Y. Shin,"Multiresolution motion analysis with applications," *The International Workshop and Human Modeling and Animation, Seoul*, 2000, pp. 131–143.

[7] P. Maillot,"Using quaternions for coding 3d transformations", *Graphics Gems I*, Academic Press Inc., Boston (1990), pp. 498–515.

[8] L. Dorst, "The inner products of geometric algebra", in *Applications of Geometric Algebra in Computer Science and Engineering,* Dorst, Doran, Lasenby(Eds.), Birkhauser, Basel, 2002, pp. 34–46

[9] A.K. Jain and J.R. Jain, "Partial differential equations and finite difference methods in image processing—part II: image restoration", *IEEE Trans. Automat.* Control 23 (1978) (5), pp. 817–834

[10] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal. Mach. Intel.* 12 (1990) (7).

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 398-401

The Research of Confidential Communication Based on the Elliptic Curve and the Combined Chaotic Mapping

Weikun Zheng, and Dongying Liang
Shenzhen Institute of Information Technology,
Shenzhen 518029, P.R. China
zhengwk@sziit.edu.cn, liangdy@sziit.com.cn

Abstract—**This paper presents a new type of chaotic encryption system based on combined chaotic mapping pseudo-random number generator, Hash Table, and elliptic curve. In this program, the elliptic curve algorithm is used for the key distribution. After the linear transformation, the original chaotic sequence generated by drive system can be combined to chaotic mapping, converted to an encryption key sequence and constructed as Hash table for message authentication. Communication experiment shows that this program is safe and is easy to implemented by software.**

Index Terms—**Hash Algorithm; elliptic curve; chaotic map**

I. INTRODUCTION

One of the hot issues in the field of misalignment scientific research in recent years is applying the chaos synchronized theory in the privacy communication system. At present, the study of chaos synchronization mainly contains complete synchronization, generalized synchronization, projection synchronization, etc. Recent literature [1] proposed the method of using a single-drive variable to set up the generalized chaos projection synchronization.

While the chaotic communication proposal mentioned above may have the following problems when chaos encrypting: In the case of low-dimensional chaos, during the short cycle caused by discretization, it is easy to find the master key while reconstructing the original chaotic dynamics from the output of low-dimensional chaos. The confidentiality of this case is not high and the information hidden in the chaotic carrier, even in the hyper-chaotic carrier, may be deciphered [2-5].This problem can be improved by high-dimensional chaos or spatiotemporal chaos. However in chaos coded communication, the amount of data transmission will increase because of the driving signal with scrambled text transmission, and high precision. The aggressor can use the known-plaintext attack to break almost all of the chaotic cryptographic system if knowing dynamics and definite and scrambled text [6] Taking into account the inherent randomness of the chaotic motion of decision It is suitable for the production of pseudo-random number using the inherent randomness of the chaotic motion, so in recent years, the research on the pseudo-random number generator based on chaos becomes a hot issue. This article will discuss how to use the combined chaotic mapping pseudo-random number generator, hash table and elliptic curve to improve the chaotic encryption proposal. The key distribution algorithm for elliptic curve will be used to ensure that the generated pseudo-random sequence retaining the properties of chaotic secure by a large extent. The pseudo-random number generated by combination of chaotic map distributes more evenly and can overcome the problem of traditional pseudo-random number generator with weak keys, which can not be simply ruled out.

II. THE NEW RANDOM NUMBER GENERATOR AND THE HASH TABLE

Due to the continuous accumulation of the limited rounding accuracy error, the traditional chaotic pseudo-random number generator (CPRNG) leads to serious doubts about the safety. The literature [12] confirmed that the security of pseudo-random sequence generator and chaotic stream cipher based on the fact that it has the clear rules of information disclosure. In this paper, for a special kind of the z-logistic chaotic map [7], a new type of random number generator is constructed according to this kind of mapping's equal definition form on rational number support territory. It can meet the precious orbit of the initial value in a limited operation precision and overcome the accumulation of errors [8]. This method can avoid the problems brought about by traditional methods and ensure the safety of CPRNG theory. Therefore, a new proposal is proposed, which is using the combination method of z-logistic chaotic map and the chaotic map with infinite collapses to improve the sensitive characteristic to the initial value and to generate uniformly distributed pseudo-random number.

The iterative of chaotic map with infinite collapses is as following:

$$x_{n+1} = f(x_n) = \mathrm{mod}(y / x_n, a) \tag{1}$$

where, $x_n \in (0, a)$, $y \in [1, \infty]$, $a \in (0,1]$, $n = 0,1,\ldots$,

When a teds to zero or y tends to infinity, this chaotic map tends to uniform distribution. And when $a = 1$, $y > 10$ or $y = 1$, $a < 0.5$, sequence can be tested through uniform statistics. Thus, if pseudo-random sequence which distributes uniformly in the range (0,1) is needed, we should assign $a = 1$, $y > 10$; that is, the situation in this paper.

AP-PROC-CS-09CN005

The z-logistic chaotic dynamics output function is as following:

$$x_{n+1} = f(x_n) = \sin(z \arcsin \sqrt{x_n}) \quad (2)$$

$$x_n = g(x_n) = \begin{cases} 0 & (x_n < 0.5) \\ 1 & (x_n > 0.5) \end{cases} \quad (3)$$

where z is parameter for the chaotic map, as is the even-number.

This CPRNG algorithm is described as below :

Step (1): The key is (h, z, t_0), h is the prime number, Z_h is defined as the collection of all integers coprimed with h. z is a generator of Z_h, h is an integer such that $1 \le t \le h-1$.

Step (2): Perform repeatedly the following actions while chaotic iteration:

$$t_n = zt_n - 1 \pmod{h} \quad (4)$$

$$x_n = \sin^2(zt_n \pi / h) \quad (5)$$

Step (3): Output pseudo-random sequence (X_n) according to the formula (3) .

Under the limited accuracy, the actual type of the chaotic orbit (X_n') generated by formula (4), formula (5) and the real orbit (X_n) meet the following results:

$$| X_n - X'_n | < \delta | \quad (6)$$

δ is the largest truncation error. When the calculation accuracy is high enough, δ tends to 0, At this time the chaotic orbit (X_n') and the real orbit (X_n) are almost identical.

A. The combined chaotic mapping pseudo-random number generator

The iterative of pseudo-random number generator which combines z-logistic chaotic map with the chaotic map with infinite collapses is as following:

$$x_{n+1} = \mathrm{mod}(y / x_n + \sin(z \arcsin \sqrt{x_n}), a) \quad (7)$$

In which, if $a = 1$ and $y > 10$, then we can get a pseudo-random number uniformly distributed between (0,1). In theory, $\{x_n\}$ can produce an ideal source of information.

The generator above demands the initial value specially, so we can make the following change:

The generator above demands the initial value specially, so we can make the following change:

$$\begin{cases} h_{n+1} = \lambda h_n (1 - h_n) \\ G = \mathrm{mod}(y / x_n + \sin(z \arcsin \sqrt{x_n}), a) + h_{n+1} \\ x_{n+1} = \begin{cases} G - 1, & G > 1 \\ G, & G \le 1 \end{cases} \\ h_n = x_n \end{cases} \quad (8)$$

In which, g is for the intermediate variable. The range of $\mathrm{mod}(y / x_n + \sin(z \arcsin \sqrt{x_n}), a)$ is [0,1), and the range of sequence $\{h_n\}$ is (0,1), and that of output sequence $\{X_n\}$ is (0,1]. These can ensure the generated pseudo-random sequence has a good uniformity although the parameter y's values are different.

B. Hash table

In the literature [9, 10, 11], W.K.Wong and others suggest updating the technology of looking-up table based on the M.S.Baptista's code to enhance safety and access to a faster encryption speed. During the encryption and decryption, the dynamic maintenance table is introduced to establish the Hash table in order to achieve the message authentication.

Before encrypting the *(K+1)th* definite text M_{k+1}, the following formula can be used to exchange the first i and j elements in the mapping table.

$$w = \left(\frac{x - x_{\min}}{x_{\max} - x_{\min}} \right) M \quad (9)$$

$$j = (i + w) \bmod M \quad (10)$$

In which M is the number of elements in mapping table and i is the table item corresponding to the definite text. Through formula (9) and (10) we can get that the maintenance of mapping table is relevant to that of getting from each encryption. So we can get Hash table used for the message authentication if encrypting all of definite text M_k.

C. The key distribution based on elliptic curve

Elliptic Curve Cryptography（ECC）uses Abel group composed of the rational points in elliptic curve and the difficulty of its discrete logarithm to build a public-key cryptosystem. The main features of ECC are high security, simple achievement and high speed. The security of elliptic curve cryptosystem is based on the difficulty of how to determine k by kP and P, which is called elliptic curve logarithm problem. At present, the relatively good methods for this problem are Pollard rho method and Pohlig-Hellman method. While the order of elliptic curves with large prime factor, these two methods will cost the time of index.

The elliptic curve on the finite field $GF(2^m)$ is composed of 2^m elements and addition and multiplication defined in the polynomial. Given m, then we can use the variables, cubic equation with its coefficients in $GF(2^m)$ and the corresponding rules of arithmetic operation to calculate.

The cubic equation of elliptic curve on $GF(2^m)$ on is:

$$y^2 + xy = x^3 + ax + b \quad (11)$$

where x and y are variables, Coefficients a and b are the elements of $GF(2^m)$, in which all of calculation will be operated.

Similarly, the point set $E_{2^m}(a,b)$ is defined as all of the integrals for formula(7).

The following can be proved that if $b \neq 0$, a limited group of Abel can be defined based on $E_{2^m}(a,b)$.

The elliptic curve on $GF(2^m)$ has higher bit utilization compared to the one on Zp.

The method of using elliptic curve to realize encryption / decryption is as following:

Each user selects at random a large integer k with the same order as q, and generates cipher text C_m, which is a couple of points:

$$C_m = \{kG, P_m + kP_B\} \tag{12}$$

The user B if wants to decrypt the cipher text, he need to use the second point minus the product of the first point and the private key of the plot B:

$$P_m + kP_B - n_B(kG) = P_m + k(n_B G) - n_B(kG) = P_m \tag{13}$$

While coding the definite text, we need respectively map data *0, 1, …, 2^m-1*, to the first 2^m elements in the set $E_{2^m}(a,b)$ and transform definite text M into elliptic curve point P_m .

D. Chaotic encryption communication program

A communication program of chaotic encryption and decryption is constructed by the above-mentioned method as figure 1, which consists of the drive system of the sending end, the response system of the receiving end, the two CPRNG systems and computer networks.

In order to ensure the correct decryption, p and r should be encrypted as two definite texts with the other ones. After the decryption of each definite text, mapping table should be transformed p times with the same encryption method, otherwise, the decryption backwards will be a wrong message

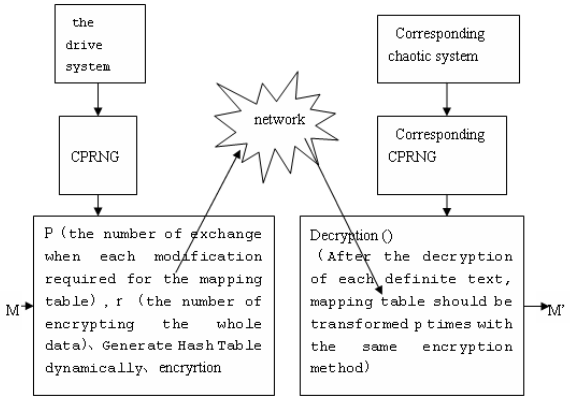


Figure1. The diagram of chaotic encryption and decryption

E. Image password Communications

Here the programming language VC6 is used to design encryption communication program. The whole process is accomplished under Windows platform.

In this case the original communication information is a landscape image (150 pixels × 113 pixels × 24 bit depth), whose effect of encrypting communication is as Figure 2. Figure 2(a) is the original image; Figure 2(b) is the encrypted image by the above program; Figure2(c) is the decrypted image by software; Figure2(d)is the decrypted image not as same as the correct parameters p and r. As can be seen that above-mentioned program is very sensitive to the parameters p and r. When there are differences in parameters, the cipher text can not be decrypted. Only if the parameters are exactly same, that can be decrypted. This is a further validation that this program has high security and good anti-deciphering characteristic.

III. CONCLUSIONS

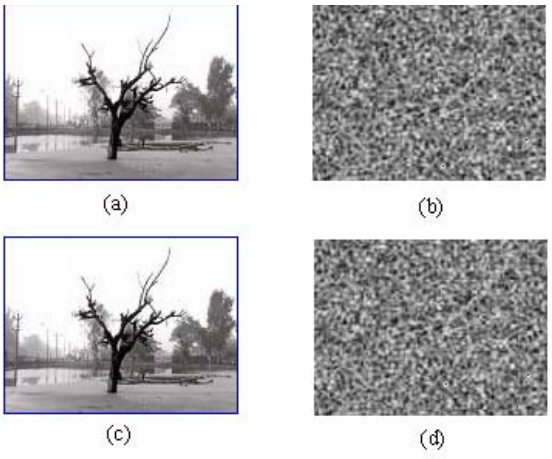


Figure2. The communication effect of static images password

This article discusses how to use the key distribution based on elliptic curve to realize the combined chaotic mapping secure communication system. Compare with the existing simple chaotic synchronization system proposed in the other literatures, the elliptic curve key distribution algorithm used in this program can ensure that the generated pseudo-random sequence by a large extent retains the properties of chaotic secure and pseudo-random number generated by the combined chaotic mapping distribute more evenly. Thus it can overcome the problem that there is a weak key in the traditional pseudo-random number generator, which can not be simply ruled out. The communication experiment used in the paper proves that the combination of combined chaotic encryption and conventional encryption is safe, feasible, and easy to implement by software.

ACKNOWLEDGMENT

The research work reported in this article has been supported in part by a Youth Scientific Funds of Shenzhen Institute of Technology (Grant No. QN-08008),and a Nature Scientific Funds of Shenzhen Institute of Technology (Grant No. LG-08004).

REFERENCES

[1] N. Li and J. Li, “Generalized projective synchronization of chaotic system based on a single driving variable and it's application in secure communication”, Acta Physica Sinica., Vol. 57, no. 10, pp. 6093-6097, 2008.

[2] J. Kuang, K. Deng and R. Huang, “An encryption approach to digital communication by using spatio-temporal chaos synchronization”, Acta Physica Sinica., Vol. 50, no. 10, pp. 1856-1860, Oct. 2001.

[3] N. Sharma and P. Poonacha, “Tracking of synchronized chaotic systems with applications to communications” Phys . Rev. E, Vol. 56, pp. 1242 – 1245, 1997.
[4] C. Zhou and C. Lai, “Extracting messages masked by chaotic signals of time-delay systems” Phys . Rev. E, Vol. 60, pp. 320 – 323, 1999.
[5] K. Short and A. Parker, “Unmasking a hyperchaotic communication scheme” Phys . Rev. E, Vol.58, pp. 1159 – 1162, 1998.
[6] J. Hu and J. Guo, “Breaking a chaotic direct sequence spreading spectrum secure communication system”, Vol, 57, pp. 1477-1484, 2008.
[7] J. González, L. Reyes, J. Suárez et al, “ Chaos-induced true randomness”, Physica A, Vol. 316, pp. 259-288, Dec 2002.
[8] S. Kawamoto and T. Horiuchi, “Algorithm for exaxt long time chaotic series and its application to cryptosystems” Int . J . Bifurc. Chaos, Vol. 14, no. 10, pp. 3607-3611, 2004.
[9] K. Wong, S. Ho and C. Yung, “A chaotic cryptography scheme for generating short ciphertext” .Physics Letters A, Vol. 310, pp. 67-73, April 2003.
[10] K. Wong, “A combined chaotic cryptographic and hashing scheme”, Physics Letters A, Vol. 307, pp. 292-298, Feb 2003.
[11] K. Wong, “A fast chaotic cryptographic scheme with dynamic lookup table”, Physics Letters A, 2002.
[12] Y. Yang and C. Jin, “Known-plaintext Attack on Chaotic Pseudo-random Sequence Generator”, Computer Engineering, Vol. 33, pp.146-148, 2007.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 402-405

Design and Implementation of Multi-Serial Ports Expansion Based on ARM Embedded Linux

Yunmi Fu [1], Yiqin Lu [1], Yanhui Zeng [1], and Bin Liu [2]
[1]School of Electronic and Information Engineering, South China University of Technology, GuangZhou, China
Email: {fuyunmi@gmail.com, eeyqlu@scut.edu.cn, yhzeng@scut.edu.cn}
[2]South China Household Electric Appliances Research Center, Shun De, China
Email: liub@hnjdy.com

***Abstract*—With the widely use of communication and intelligent devices, more and more extension modules are attached to an arm embedded system, most of them are through serial port. Thus to extend the serial ports of an arm system is necessary. This paper proposes a method of serial port expansion based on SC16C554. With SC16C554, system bus is used to extend four serial ports, which have the standard modem interface and work independently. Details of working principle of hardware and design method of device drivers of serial port expansion are presented.**

***Index Terms*—SC16C554, S3C2410A, Serial Port Expansion, Device Driver**

I. INTRODUCTION

With the increasing requirement of digital home and intelligent industrial control, embedded systems need to increase the various peripheral modules, such as the GPS module, GPRS/GSM module, ZIGBEE module, X10 module, etc [7]. These modules usually communicate with the CPU by serial ports. Because the serial interface devices have the advantages: flexible control, simple interface, occupying less resource. So they are widely used in Industrial Control, Smart Home and Prevention Technology areas [12]. But the ARM microprocessors usually provide limited serial ports, 2 or 3. One serial port is usually used for PC control, and there are only one or two serial ports to use to communicate with the microprocessor for slave devices. If a home gateway system needs to use some serial devices at the same time, such as GPS module, GPRS/GSM module, ZIGBEE module, X10 module, this home gateway system will need more serial ports resource. Therefore, the embedded system with multi-serial port expansion can be an effective solution to this problem.

II. SERIAL PORT EXTENSION METHODS

For the serial port of embedded system in the problem of insufficient, here are several common methods for serial port expansion: Software Simulation Method (SSM), Serial Port Extend Serial Port Method (SPESPM), USB Port Extend Serial Port Method (UPESM), Ethernet Port Extend Serial Port Method (EPESM) and Parallel Port Extend Serial Port Method (PPESM).

A. Software Simulation Method

SSM is based on transmission formats of serial port communication, using the timers and I/O ports of the host to simulate the serial port communication timing, in order to achieve the purpose of extending serial port. Advantages are low cost, but the reliability is poor and development of software is very difficult.

B. Serial Port Extend Serial Port Method

SPESPM usually uses the extension chip with a choice of address to extend the serial port or uses the software control timing to extend serial ports. There are two main chips of GM8123/25 and SP2338 series to use to extend the serial port base on serial to serial [5]. Advantages are simple control, occupying less resources, versatility and good stability. But the communication parameters can be less editorial, and the expanding serial ports can not work independently.

C. USB Port Extend Serial Port Method

UPESM is using a dedicated chip convert USB port to serial port. Advantages are plug, easy to expand and not needing additionally power supply, but the cost is very high, and the expansion of multi-serial port is also more complicated [5].

D. Ethernet Port Extend Serial Port Method

EPESM is using Ethernet interface to change to the serial port. The disadvantage of this method is high cost, and the design is more complex [5].

E. Parallel Port Extend Serial Port Method

PPESM is used SC16C554 to extend four serial ports which can work independently at the same time. The SC16C554 is a 4-channel Universal Asynchronous Receiver and Transmitter used for serial data communications. Through writing the control register LCR, IER, DLL, DLM, MCR, and FCR, it can achieve SC16C554 serial channels communication [1]. Baud rate generator (BRG) of serial channel allows the clock to divide any number between 1 and 65535. According to their different frequencies in one of three kinds of common standards, BRG determine the baud rate. According to regulate an external crystal, SC16C554 can

This work was supported by Guangdong Foundation of Science and Technology Project (2006A10101003, 2006A1020300, 2008B090500073), Guangzhou Foundation of Science and Technology Project (2003B11609), Project of Technology Breakthrough in Key Fields of Guangdong and Hong Kong (2006Z1)

get a very quasi-baud rate and produce many different types of baud rate from 110bps to 460800bps.

Although PPESM is a little complicated to control, taking up more resources of MCU ports, such as the I/O ports, interrupt resources, but PPESM can provide MODEM control signals and expand out four serial ports which can work independently at the same time and control flexibly , communicate by high-speed and meet general serial port settings. Since we use a powerful, resource-rich S3C2410A which has 117 general-purpose I/O ports and 24-channel external interrupt source as the controlling chip, so the problem of needing many resources can be solved [2]. This paper uses this method to extend the serial ports.

III. Design Of Multi-Serial Port Expansion Interface

A. SC16C554 internal structure and working principle

The SC16C554/554D is a 4-channel Universal Asynchronous Receiver and Transmitter used for serial data communications [1]. Each channel can receive serial data from peripheral and convert them to parallel data for CPU, also can convert the parallel data from CPU to serial data and sent to the peripheral. As the interface between the CPU and the SC16C554 is based on parallel mode transmission, the interface between the SC16C554 and the peripheral is based on serial mode transmission, therefore, there must be Receive Shift Register (RSR) and Transmit Shift Register (TSR) in the block diagram of SC16C554 [1]. The SC16C554 block diagram was shown in Figure 1.

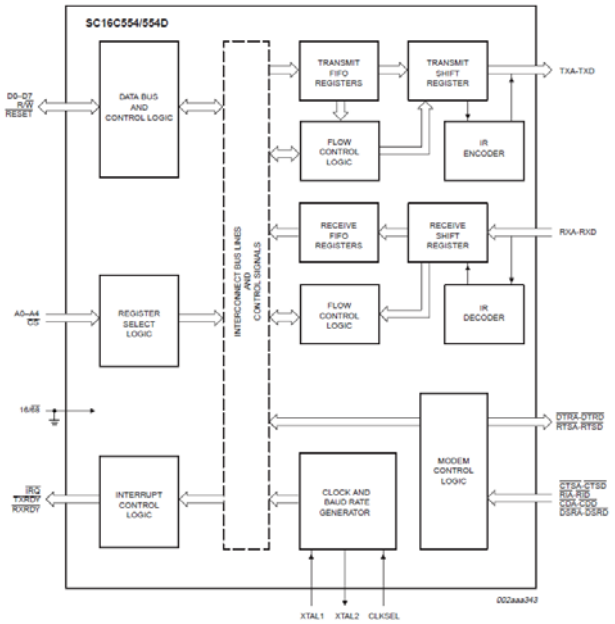


Figure 1. SC16C554/554D block diagram [1]

When the data from peripheral were sent to SC16C554, the data were sent to the RSR one bit by one bit. Once the RSR has received one byte data, the data from RSR were sent to the Receive FIFO Register (RFR). And then the CPU receives the data received by RFR. In the data output process, CPU sent the parallel data to Transmit FIFO Register (TFR), and TFR sent the received data to the TSR, then TSR convert the parallel data to serial data and sent to the peripheral one bit by one bit [12].

B. SC16C554 internal registers

Different combinations of address lines A0, A1, A2 of SC16C554 internal registers represent different registers. Table I details the assigned bit functions for the SC16C554 internal registers.

TABLE I. SC16C554 INTERNAL REGISTERS [1]

A2	A1	A0	Read Mode	Write Mode
0	0	0	RHR	THR
0	0	1	IER	IER
0	1	0	ISR	FCR
0	1	1	LCR	LCR
1	0	0	MCR	MCR
1	0	1	LSR	/
1	1	0	MSR	/
1	1	1	SPR	SPR

C. Schematic description

The circuit of Serial Port Expansion part was shown in Figure 2 and 3. SC16C554 data lines D0-D7 were connected with the CPU's bus DATA0-7. The nIOR, nIOW lines were respectively connected with the read signal nOE line and write signal new line. The High-level RESET signal of SC16C554 was connected with the

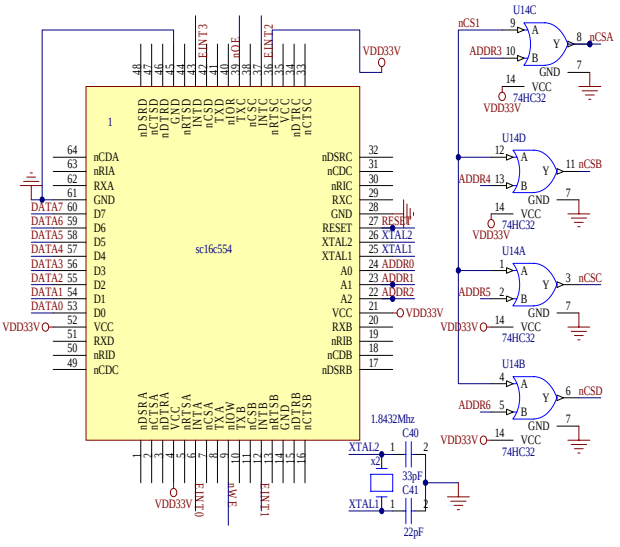


Figure 2. The circuit diagram of SC16C554

RESET signal of CPU's hardware system. Interrupt lines INTA-C of 4 serial ports were connected with the EINT0-3 of CPU respectively, which used for sending the interrupt signal to CPU when data were received or sent. There are 15 registers in each serial port and there are seven registers are re-used. Address lines A0-2 of register logic control were connected with the CPU's bus ADDR0-2 respectively. The Chip Selection (CS) signal nCSA-D of serial ports were independent. The right of Figure 2 is a decoding circuit of low-level CS. Through the CPU-nCS1 Logical OR the ADDR3-6, we can get the base address of 4 serial ports expansion respectively is 0x08000070, 0x08000068, 0x08000058 and 0x08000038 [1,8].

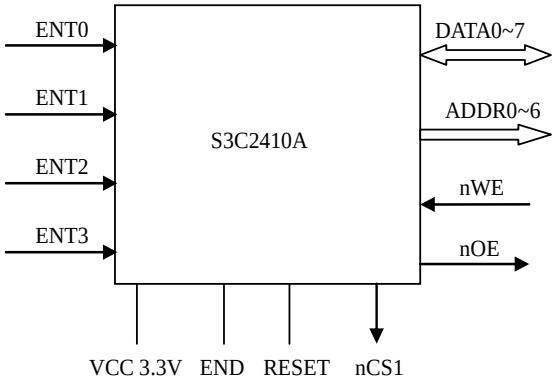


Figure 3. Serial Extension connection diagram of S3C2410A

IV. Driver Design

A. Initialization of S3C2410A

S3C2410A must be initialized to complete to set the parameter of SC16C554, before the system work [3]. It mainly include: Register the device to the system, Mapped the device address to the virtual address, Set BWSCON, BANKCON Reg to set the timing, Apply for interrupts resources and interrupt mode from system, Set the various expansion serial port communication baud rate, data frame of data bits, stop bits and parity, etc [3, 4, 10].

It was shown at the following initialization function EXT_COM _init(). Note: This system uses the embedded Linux 2.6.11 kernel.

```
static int __init EXT_COM_init(void)
{
   ……
   ret=register_chrdev(EXT_COM_MAJOR,DEVICE
   _NAME,& EXT_COM_fops);
    // Register the device to the system
for(i=0;i<4;i++)
{
   vEXT_COM_ADDR[i]=ioremap(pEXT_COM_AD
   DR[i],16); // Device is mapped to the virtual address
}
__raw_writel((__raw_readl(S3C2410_BWSCON))&~(
0xf<<4),S3C2410_BWSCON );
  //According to use the nCS1,  Set the BANKCON1 to
    set the timing [2]
__raw_writel(0x1f4c,S3C2410_BANKCON1);
for(i=0;i<4;i++)
{
   EXT_COM_Init(vEXT_COM_ADDR[i],EXT_CO
   M_PARAM[i]);
    //Through the function to set the serial port
      communications of the baud rate, data frame of
      data bits, stop bits and parity parameters
   request_irq(EXT_COM_INT[i],uart_irq_handle,SA
   _INTERRUPT,DEVICE_NAME,NULL;
    // Apply Interrupt Resources
   set_irq_type(EXT_COM_INT[i],IRQT_HIGH);
    // Set the interrupt mode to the high-level trigger
}
……
return(0);
}
```

B. Driver file_operation function description

Several operational functions used in this design are as follows [6]:

```
static int EXT_COM_open(struct inode *inode,struct
    file *file)  {…};// Open device
static int EXT_COM_release (struct inode
    *inode,struct file *file) {…} ; // Release resources
static ssize_t EXT_COM_read(struct file *file,char
   *buf,size_t count,loff_t *f_pos) {…}; //Read device
static ssize_t EXT_COM_write(struct file,const char
   *buf,size_t count,loff_t *f_pos) {…};//Write device
static int EXT_COM_ioctl(struct inode *inode, struct
   file *file,unsigned int cmd,unsigned long arg) {…};
   //set  the serial communication of baud rate, data bits,
   stop   bits and parity parameters
static irqreturn_t EXT_COM_irq_handle(int irq,void
   *dev_id, struct pt_regs *regs);  // Interrupt handler
```

Due to space limited, the upper functions are not detail.

V. Experiment Result

This system uses the embedded Linux 2.6.11 kernel.

Test device driver process steps: (1) Load the module, (2) Build the device nodes, (3) Run the program for testing [9].

Run “ #insmod sc16c554.ko ” to load the module and run “#rmmod sc16c554.ko” to unload the module. Run “ #mknod /dev/EXTCOM0 c 233 0 ” to add the first serial port node of expansion. “ mknod ” command is use to add a node. The first serial port node named EXTCOM0. 'C' represents a character device, and 'B' represents a block device. The sc16c554 is a character device, so here use 'C'. The ‘233’ represent the number of major device, and ‘0’ represent the number of minor device [11].

The following is a test application:

In this design, communication parameters of expansion serial ports can be edited, such as baud rate, data bits, stop bits, parity efficacy, flow control, etc. The following is used 9600 bit/s baud rate for the test: (The test program set the communication parameters: Baud Rate: 9600 bit/s, data bits: 8 bit, Stop bit: 1 bit, parity checksum: none, Flow Control: None [13].)

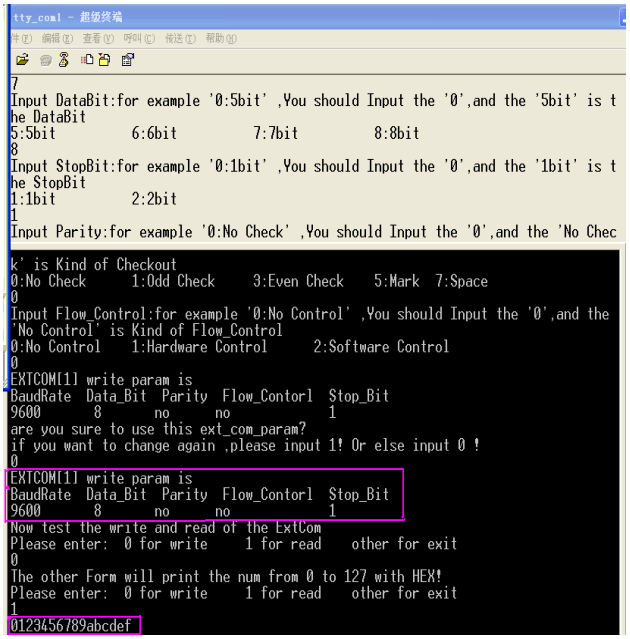


Figure 4. Test program running results

Figure 5. Serial Test Window

Run “#./test ” to test the driver, the test is a program file. The running results were shown in Figure 4 and Figure 5.

VI. Conclusions

In this paper, expansion chip of SC16C554 uses system bus to extend out 4 serial ports which can work independently. SC16C554 uses interrupt work mode to improve the efficiency of ARM9 system, solving the problem of inadequate serial port of embedded systems. If you need to add more expansion serial ports than four, you need one additional SC16C554 chip. Two SC16C554 chips can be expanded to eight serial ports. Using this method, the system can work steadily. Experiments proved to be a viable method of multi-serial ports expansion of embedded system.

References

[1] Koninklijke Philips Electronics.SC16C554/554D Datasheet[Z].Rev.01.2002

[2] Samsung Electronics.S3C2410 32-BIT RISC MICROPROCESSOR USER'S MANUAL[Z]. Revision 1.2, 2003

[3] MA Zhongmei, LI Shanping, KANG Kai, ARM & Linux Embedded Systems Tutorial, Beijing University of Aeronautics & Astronautics Press, Beijing,2004

[4] Jonathan Corbet, Alessandro Rubini, Geg Kroah-hartman, Linux Device Drivers 3rd Edition,O'reilly, 2005

[5] Song Guomin, Study and Implementation of Cost-Effective Multi-Serial Port Expansion, Chengdu Electromechanical College, 2005

[6] Sun Tianze, YUAN Wenju, ZHANG Haifeng, Embedded Design and Linux Driver Developer's Guide, Electronics Industry Press,2005

[7] Lan Lirong, Design and Development of Embedded Home Gateway Hardware System for NGN, South China University of Technology Press, 2007

[8] Samsung Electronics.K9F1208U0M-YIB0 FLASH MEMORY Datasheet[Z] .Revision 1.8, 2001

[9] FriendlyARM, Embedded development board QQ2440 User Manual, Guangzhou FriendlyArm Computer Technology Co., Ltd, 2007

[10] Rubini, A., Corbet, J.: Linux Device Drivers. Ed. O'Reilly, 2001

[11] A Nan, Study Notes of Introduction to Embedded Linux, http://bbs.21ic.com, 2002

[12] Ding Jidong, Design of Serial Port Expansion and Asynchronous Communication Interface Based on S3C44B0X and TL16C554, Industrial Control Computer, 2005,18(3):21,73.

[13] Prolific. PL – 2303 USB to RS - 232 Bridge Cont roller V1. 4 ,2002. 8.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 406-409

Study of Visual Object Tracking Technique

Yumei Xiong [1,2], and Yiming Chen [2]
[1] Electronic & Information School, Shanghai Dianji Uneversity, Shanghai , China
e-mail:ymperi@sina.com
[2]School of Computer Engineering and Science, Shanghai University, Shanghai , China
e-mail: jszlee@sina.com

***Abstract*—Analyzing the traditional visual object recognition algorithm, we presented a complete set of visual target recognition and object tracking algorithm. The algorithm uses image processing method to calculate the frame and characteristic points. And a predictable based on the "search window" approach was given in the algorithm. Experiments have shown that this target tracking algorithm speed up processing speed, the algorithm can provide excellent tracking and detection, and real-time processing.**

***Index Terms*—tracking, search-window, characteristic point, visual object**

I. INTRODUCTION

Visual target recognition and tracking is to identify the location of visual signs, to calculate the interested target state from sequence of images, Then using different tracking strategy according to target nature, degree of freedom and tracking conditions[1,2]. Target tracking technique requires a better recognition rate and high real-time [3,4].

In the current application, target is tracked by video collected by CL, CR camera. And calculating location of the real object, then the virtual objects will be placed in this position through a series of space coordinate conversion. In addition, the target tracking technique can be applied to the image captured by VL, VR camera too. The image of virtual objects can be directly added to the sign's location collected by video image. This approach is simple and fast and has a good effect of augmented reality. As it can be seen, target tracking technology is key technologies of the augmented reality system, and is important to the expansion of system functions.

In this paper, visual target identification and target tracking technology were studied in detail. And in tracking mode, the paper presents a predictable approach based on the "search window", which speed up processing speed of the target tracking algorithm.

II. VISUAL OBJECT RECOGNITION

In the recognition of visual signs, the general information that can be used is the edge information, geometric information, color information [5]. Our system has no rich colors, so identification using only the edge information and geometric information [6]. We designed the visual target recognition algorithm. This identification process can be divided into three steps: the first step will be collected from the image into a binary image (black and white images).The second step requires binary image calculated based on the signs surrounded by boxes; third step is looking for a white box surrounded by a circular feature points.

A. The original image is converted to binary image

In order to improve the algorithm real-time performance, we first thought of the method transferring image into a binary image. In the color image into a binary image of the process, it was divided into a two-step, the first step the color image into a 256 gray-scale image; second step used in image processing threshold segmentation algorithm, the gray-scale image divided into two binary images.

Color image into a gray image as follows:

Step 1. Matrox capture card to acquire images through RGB format color image, each pixel is used R, G, B three bytes of storage.

Step 2. The RGB color system image to YUV color system, in which Y represents the pixel brightness signal;

Step 3. Y signal directly out of each pixel is a representative of the gray level 256.

Transferring gray image into a black and white binary images, this step can be highlighted in the image visual signs. This process is commonly used threshold segmentation method. Threshold segmentation is an image segmentation technique. Its purpose is to divide image space to some meaningful regions [7].

One of the most key point of the threshold segmentation lies in how best to choose the threshold. As the scenery and the light may change, we can not simply determine the threshold. We use the minimum error segmentation method. The basic idea is to find a threshold, according to which, error probability of division the target and background is the smallest. Referring the idea of the minimum error segmentation, we use the iterative method to realize an optimal segmentation threshold algorithm. Algorithm steps are as follows:

Step 1. Find the image minimum and maximum gray value of Z_{min} and Z_{max} so that the threshold $T^0 = (Z_{min} + Z_{max}) / 2$.

Step 2. According to the threshold Tk image divided into two parts, the object and background, find the average gray value of two-part Z_{low} and Z_{high}

$$Z_{low} = \frac{\sum z(i,j) \times N(i,j)}{\sum_{z(i,j)<T^k} N(i,j)}, \quad Z_{high} = \frac{\sum z(i,j) \times N(i,j)}{\sum_{z(i,j)>T^k} N(i,j)}$$

Supported by 863 project（2007AA01Z319）

Z (i, j) and N (i, j) are the image point (i, j) the gray value and the right to make the calculation of N (i, j) = 1; 3. Find a new threshold value: $T^{k+1} = (Z_{low} + Z_{high}) / 24$. If $T^k = T^{k+1}$, then the end, or else K ← K +1, go to step 2.

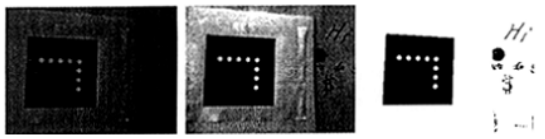


Figure 1. Image is converted into binary image.

After transferring the RGB color image to 256-color grayscale and the gray level threshold segmentation algorithm, the input image was successfully converted into binary image. Shown in Figure 1, the left is a original image, the middle grayscale image is converted from the original image, while the right is binary image obtained after the gray-scale images using the optimal threshold segmentation.

B. Calculating the sign surrounded by boxes

We can more quickly calculate the visual signs of the image information. We finally need the coordinates of eight feature points of the sign. We consider the search in two steps: first detection of signs of the frame, and then search in the frame of the regions of feature points. Frame detection involves three steps: to eliminate noise, extract and calculation of border coordinates.

Figure 2. Corrosion de-noising.

Step 1: Using image corrosion to eliminate noise of binary image[7]. The result is elimination of corrosion of the frame connected region. When we carried out corrosion of binary images, all black areas are properly connected smaller, some small regions are eliminated, The result is shown in Figure 2, black-connected region become smaller or lost, which indicated it has a good de-noising performance.

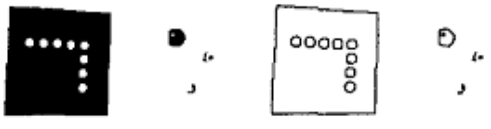


Figure 3. Edge detection.

Step 2: Extract sign border. We use the contour extraction to hollow out of internal points. a new image S1 generated after image S is corroded. S1's change is as in Figure 3. After this step, closed-connected region disappeared, the image leaves a very small black spots. This provides the convenience of straight-line fitting.

Step 3: Calculate the coordinates of the rectangular border. We use straight-line fitting method to analyze the image of all points, first constructed some of these points may fit a straight line, and then find the four closest to the quadrilateral geometric restrictions on the border line, according to derive the intersection of these four linear frame surrounded by the visual signs . Straight-line fitting method, Hough transform run through several times more than fitting a straight line, we have based on the principle of Hough change to improve the Hough transform procedure to detect the image of the quadrilateral area, the algorithm is as follows:

Step 1. Initialize a transform domain r, θ space array, r direction to quantify the number of pixels for the image diagonal direction, θ the direction of quantifying the number of 90 (angle from 0-180, each cell 2 degrees) ;

Step 2. Sequential search in the image all black spots, for each black point, the corresponding points in the transformation add 1;

Step 3. Find the transform domain maximum records;

step 4. Will be near the point of maximum points and cleared;

Step 5. Find the maximum value corresponding to a straight line, and store it, if the total of the six straight focus has turned to six, or turn 5;

Step 6. The use of quadrilateral geometric constraint relations, the six candidates in a straight line to find the nearest four straight-line projection of the visual signs that this is a marker of the frame;

Step 7. According to this calculation of four straight lines mark the four vertices.

One of the quadrilateral geometric constraints, simply put, that is similar to the slope of a straight line, a maximum distance of the two elections; for the slope of a straight line related to large, select them more freedom. Finally make sure you select four straight, basically two sets of parallel lines, a larger angle between the two groups of equal lines, and the similarity between the distance apart.

C. Solving a collection of feature points

The identification of visual signs provide the coordinates of projection of feature point in different angle, and the classification and recognition in the visual signs also make use of these feature points, so scanning the visual signs need to find feature points finally. We designed a solving the feature points of the algorithm, which is described as follows:

Loop detecting each pixel point in the sign surrounded of box; if it is a white point, and it is Surrounded by four straight lines in the frame. The seed filling algorithm is called to fill the connectivity of regions where it is in. In the filling process, it will count in the region pixel number (area), the number of boundary pixels (perimeter), surrounded by boxes, and centroid. If the general centroid is located in the center surrounded by boxes, then the calculation of its degree at the regional round(that is with the square of circumference divided by the area). If the value is nearly 4π, and that the area of the region is less than a certain threshold value, then that the region is the projection of

the characteristic point, the centroid and the region will be recorded to the collection of feature points.

Feature point detection of the input images have not yet carried out by the corrosion of binary image obtained at the sign box surrounded conducted on this sub-graph for seed filling, almost no interference by other white areas, so a higher success rate. But the point does not rule out the possibility of interference is detected, seeds and fill out all feature points detected will be followed by them as a candidate feature points, through the signs of basic geometric arrangement of rules limiting the lattice, to select the appropriate feature points. We need to be detected by eight feature points for the visual symbol identification and tracking. Suppose the set of all candidate feature points for the S, S a total of n elements, then the candidate feature points from the set of selected characteristics of the final eight points algorithm is as follows:

Step 1. First of all pairs of feature points S in the n-region to scan, statistics of their area, an area classified as a small gap between the classes. If a class size of less than eight times the value of occurrences out of it, removed from the collection to remove remaining after the S in the n 'elements;

step 2. Remove S feature points of all n-element centroid coordinates for Hough bad, the probability of fitting out the two most likely on the line, and calculate the warring A;

step 3. Remove S with the smallest distance from point A to point A ', from A' departure, if a straight line along the two were able to search a certain direction to four and three characteristic points, then the search success, clear S, these seven features point and A 'point into the set S; otherwise search fails, do not remove S.

So far, we have completed the entire visual target recognition algorithms, the algorithm will be collected by the image into a binary image to calculate the symbol surrounded by boxes and feature points, in order to prepare for the back tracking.

III. VISUAL SIGNS PREDICTION AND TRACKING

Prediction Tracking is used to track the visual signs of movement, and predicted the emergence of its next position, which is to improve the real-time nature of the necessary means. The traditional calculation method for computer vision is poor in real time. Our response is: In the visual symbol recognition, visual signs bearing a square frame, this frame as a "search window", each time only "search window" within the image for processing. "Search window" moves as the visual sign' movement to ensure that the visual signs every fall in this range. Of course, the "search window" coordinate system image coordinate system have a certain deviation between the calculation results must be converted through the shift can be passed to the three-dimensional reconstruction of blocks, but this time in terms of complexity of the computer is almost negligible. As a result, the method does not affect the three-dimensional reconstruction of the situation, reduces the complexity of image processing.

Tracking the most crucial is the need to get a sign in front of the projected coordinates to predict the next symbol will appear, and calculate the next "search window" [8]. Forecasting methods we use relatively simple, that is, the last two results, based on the inertia of the rules to determine the use of objects predicted the next area, and an appropriate amplification in the region.

In ideal circumstances, each can be detected by visual signs, based on the "search window" of the target tracking algorithm is described as follows:

Step 1.The initial conditions of zero default processing the first frame, the frame of the "search window" and the entire image the same size.

Step 2.In the first frame when the "search window" and the entire image the same size. On the "search window" in the image recognition algorithm is called the static visual symbol identification. Repeatedly adjusted to ensure that a simpler context, the first frame to successfully identify, access to sign an appropriate set of feature points surrounded by boxes;

Step 3.In the second frame after every time he returns to see the former two "search window" of four vertices, according to which two vertices of room to move between groups is projected third "search window" of four vertices the possible direction of movement, and the four directions in accordance with this amendment "search window" of size and region, as the expansion / shrink / move the strategy;

In less than ideal circumstances can not be detected by visual signs, tracking algorithms need to consider: If the last one detected characteristic points less than the actual number, then the "search window" to the four directions of expansion of a certain distance; the other hand reduce the . Reality is achieved, to expand and contract the distance of feature points, the average side length of the external box.

The application on the "search window" tracking algorithm reduced complexity image processing, improve processing speed, to provide the possibility of real-time visual 3D tracking sign-up.

IV. TEST RESULTS

In the actual operating environment, we use two high-speed computer: Dual XEON 2G processor, 2G RAMBUS memory, 80G 15,000 transfer SCSI hard drives, Matrox Meteor II image capture card. And two computers connected via Gigabit LAN. In order to measure the performance of target tracking algorithms, we tested from the success rate of signs and search speed. The success rate of the signs is the number of frames to detected visual signs of the percentage of all frames. Search speed refers to that the algorithm can handle the number of frames in the unit of time.

We test the algorithm in external conditions similar to real-world environments. Testing is done in a unified natural light conditions and relatively simple background. Cameras take photograph backlight. And experiment lasted five minutes. Image acquisition frequency is set to 30fps. Natural light is softer lighting, no flicker. A simple background is that the background color, texture is more

uniform and have no a large number of the sudden changing point in colors. In natural light and a simple background conditions, the threshold segmentation algorithm can be better to cut the visual signs and image background. Before the test, you must adjust the camera's focal length to be able to take the more clearly image of the visual sign from a fixed distance,.

Part of the actual test procedure screenshot is as in figure 4, respectively, one is untreated image, the other is processed image of collection of feature points.

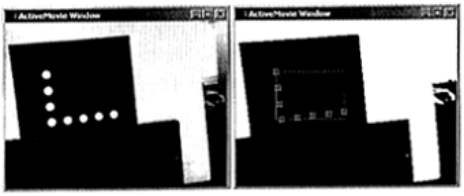


Figure 4. Test results Figure

Tests were done in three cases. The first case: adjusting the location of signs and the distance to the camera to ensure the success of the detection in the first frame of signs in order to establish the correct search window that can correctly predict the next sign occurrence, and the visual signs will not be completely ran out of camera view and move in the region with clearer images in the camera; the second case: failure to detect the first frame, the algorithm must quickly detect the location of signs and visual signs occasionally ran out of camera vision; third situation: the visual sign will move to the area imaging blur in the testing process. The test data of these three cases is as shown in Table 1.

TABLE 1. ALGORITHM TEST RESULTS

	CASE1	CASE 2	CASE 3
Success rate (%)	95.8	79.5	56.4
Speed (fps)	26.1	23.9	19.1

As can be seen, the algorithm had good search performance and real-time under searching successfully in the first frame, and no complex lighting and the background interference, and not ran out of camera view cases. When the visual signs occasionally ran out of camera view, the algorithm can quickly search the visual signs coming back to the camera view. Although the success rate dropped to eighty percent, there are better real-time, the success rate can be acceptable. When the visual signs are out of the scope of clear image, the searching success rate decline greatly, but the real-time can be tolerated.

V. CONCLUSION

The visual signs of identification and target tracking techniques were studied in detail, in conjunction with their application needs. Analyzing and optimizing the traditional visual object recognition algorithm, we presented a complete set of visual target recognition and object tracing algorithm. The algorithm uses image processing method to calculate the frame and characteristic points. And a predictable based on the "search window" approach was given in the algorithm. Experiments have shown that this target tracking algorithm speed up processing speed, the algorithm can provide excellent tracking and detection, and real-time processing.

ACKNOWLEDGMENT

This research was supported by 863 project （2007AA01Z319）The authors wish to thank Li Chao and Wang Xuejun in building model.

REFERENCES

[1] Zhongyang, Xu， Hao, Chen， Hui, Ding,Scheduling algorithm for MPEG-2 TS multiplexers in CATV networks,IEEE Transactions on Broadcasting, v 46, n 4, p 249-255, 2005

[1] CollinsR T, Liu Y, LeordeanuM, et a.l Online selection of discrim inative tracking features [J].IEEE Transactions on Pattern Analysis andMachine Intelligence,2005,27(10):1631-1643.

[2] Hsu D F, LyonsD M, Ai J, et a.l Feature selection for real-time tracking [C] //Proceedings of SPIE. Kissimmee, FL,USA,2006,6242:163-170.

[3] Comaniciu D, Ramesh V, Meer P. Kernel-based object tracking [J]. IEEE Trans Pattern Analysis and Machine Intelligence, 2003, 25(5): 564-575.

[4] Hou Z, Han C, Zheng L, et a.l A fastvisual tracking algorithm based on circle pixelsmatching [C].Proceedings of the Sixth InternationalConference on Information Fusion. Cairns, Australia,2003:291-295.

[5] Stan Birchfield, Elliptical Head Tracking Using Intensity Gradients and Color Histogramsm, Proceedings of CVPR, 1998, pp.232-237

[6] Ayala-Ram irezV, ParraC, DevyM. Active tracking based on Hausdorffmatching[C] .InternationlConference on Pat-tern Recognition.2000,15:706-709.

[7] Eric Foxlin， Extended draft version of Chapter of Handbook of Virtual Environment Technology[M] ， Lawrence Erlbaum Associates，2002

[8] Feng yan，Chen yimin，Rectification of magnetic force tracker using neural network in AR system[J]，Journal of shanghai university，2006，10(5)：431-435

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 410-413

Application of PTT in Digital Library

Zhonghua Deng [1] , and Youlin Zhao [2]
The Center for the Studies of Information Resources, Wuhan University, Wuhan, China
Email: [1]sim.dzh@gmail.com; [2]zhaoyoulin.1986@163.com

***Abstract*—The 3G era brings new opportunities to telecommunications industry. This article introduces current situation of using network technology in digital library, in the 3G network, PTT has its own advantages so that it is feasible to be used in digital library. We highlight instance about digital library using PTT, and then solve the issue of context-aware transformation and use SIP-based system architecture to reduce time delaying. In the last part, we analyze some issues about PTT applications in 3G network, such as tariff, mobile terminal, unused resources, and discuss corresponding solutions.**

***Index Terms*—3G, smart mobile phone, PTT, digital library**

I. INTRODUCTION

In the modern times, the 3G technology is gradually implementing in telecommunications industry, so that users will not be restricted by locations and time. As communication technology develops and the functions of movable termination enhance, the value-added services are gradually involved in people's daily life. The PTT (Push to talk) can be used in digital library through smart mobile phone, to meet users' requirements, such as online reading, online video, online class and online communication.

By Internet, computer functions are used in all aspects of people's daily life, but smart mobile phone is less multifunctional compared with computer. Though streaming media has been used in digital library, online communication has not achieved. If IVR service [1] is applied to digital library, some functions such as voice SMS, voice-enabled chat, teleconference and voice-on-demand, as well as online conference and online schooling. The 3G era makes it possible that some successful technologies in some other fields can be used in digital library, so PTT, as a typical case of IVR service, must be an expecting service.

PTT implements following functions, we call without dialing, and only need to press a special key and our voice can be transmit to any members in the group. It seems like interphone, but not need to impropriate some special channel and network. It also can be extended to all networks that 3G covered [2]. Applying PTT to digital library, some functions can be implemented as following, experts teach online, readers watch online, and the experts communicate with readers online, even communicate with other members in group. As is well-known, this function has been implemented by computer earlier, but PTT technology is used in interphone, which can be used for reference to apply PTT to digital library.

II. MOBILE SERVICE OF DIGITAL LIBRARY

3G indicates the third generation mobile communication technology, compared with the first generation called analog system and the second generation as GSM, CDMA technology [3]. As the emergence of Internet, the library resources are gradually digitized. Mobile services in digital library will develop continuously due to the constantly progressive communication technology and the constantly enriched mobile terminal function. 3G technology used by Computers can increase the speed of network access, while 3G technology used by smart phone terminal is convenient for readers to refer to digital library resources anytime and anywhere. There are two kinds of services to share digital library resources by smart mobile phone [4].

A. WEB browsing

The exchange of librarians and readers in the 3G era is more convenient. With the acceleration of speed, 3G smart phones have lots of interactive modes such as video, voice, words and so on. These modes are no longer subject to time and place restrictions.

Libraries could integrate the large sum of information resources and the current system functions into the WAP net. Readers could enjoy the resources of digital library anywhere and anytime through smart phone or other smart terminal which use 3G technology to require information.

B. Streaming media data transmission

The present smart mobile phone can browse online and upload streaming media. The files of video lecture, CD attached books, movies, phonic video are very large, but readers can browse online by smart phone under 3G network [4].

With the advent of 3G network, the field of voice communication in digital library can be strengthened. For example, PTT (push to talk) technology based on 3G network can achieve on-line communication [5]. Wireless PTT technology is a type of unidirectional speech group communication and it is one of the businesses with greater development potential. It is very similar with the interphone. Readers could watch video online and communicate between groups closely face-to-face. This can realize this function in 3G era, which cannot be achieved earlier in the 2G era

This work was supported in part by the Ministry of Education under Social Science Research Youth Project Fund, (project code: 08JC870009)

III. APPLICATION ABOUT PTT TECHNOLOGY IN DIGITAL LIBRARY

A. PTT in 3G environment

PTT technology has requirements about time to these terminals. Internet access speed of 3G mobile phone is 15-20 times quicker than 2G or 2.5G, which provides guarantee in time to realize the PTT technology. A typical example of PTT communication as follows: User A pushes down the button to talk 2 to 3 seconds and then waits for the answer from User B; B receives the information from A. Then he cerebrates for 2 to 3 seconds and pushes down the button to answer for 2 or 3 seconds. Once communication always includes two or three bidirectional procedures, so a whole communication may takes 16 seconds [6]. PTT is a simple, high-speed and flexible form of group voice communication. It's aiming at reducing the delay time in communications.

B. IPv6, Multi-User Detection, Assisted Global Positioning System technical support

3G provides a better network environment for the digital libraries. To use this environment rationally, we should allocate IP address of smart mobile phone. Address capacity of IPv6 is about 8*1028 times of IPv4 [7]. It is large enough to make every computer of the world has an IP address, and can fully meet assignment of handheld libraries' address of mobile phone in the future. In addition, with smaller routing tables, IPv6 enhances the support for group communication, and more adapt to online video and online intercom.

First of all, we should ensure the safety of information during the process of PTT service. It is prone to a lot of information redundancy, the exchange of information asymmetry, incomplete information transmission and the phenomenon of multiple-access interference in the multi-user communication during PTT service. In the CDMA system, MUD (multi-user detection) is to eliminate multiple access interference technology in the latest developments [8]. MUD ease issues of the distance effects from the base station between the user and digital library. It improves system performance and increases CDMA system capacity significantly.

A-GPS （Assisted Global Positioning System）technology can be used from 2G to 3G, which requires network and mobile terminals capable of receiving GPS information [10]. It is a technology combining of network base station information and GPS information to locate mobile terminals. Compared with other positioning technologies, A-GPS, as a high accuracy mobile location technology, has some benefits [8]: (1)Shorten the time of the initial positioning (2)Improve the receiver sensitivity, expand the coverage of Location- Based Services (3) Improve the positioning accuracy.

C. The integrity of PTT system

PTT system use client/server solution, its basic components can integrate PTT client with ready-made or the next generation of mobile phones. Multi-user can operate on digital library using PTT services through integrating a PTT server on the wireless packet network operators. PTT system is mainly composed of the following three levels [5]:

PTT server: PTT switch, which serves as a bridge between the various components in PTT system, is the core equipment of PTT. PTT server provides group call, encoding, decoding, billing interfaces, border gateway connection interface, radio gateway interface and server connection status display and so on. PTT server has good scalability.

PTT client: PTT client applies the server mode. Multi-party call is more powerful in the full-server model, which can support cluster call up to 50 people. Using the server mode ensures the number of participants and the stability of network in the digital library.

Database server: We usually adopt Oracle system, a centralized database with cluster structure, to handle all users' data. Thus, the entire users' data in the database server are correct and perfect.

D. Integrated management of three networks

In 2G, the management conducts a "single management", that is only for the management of telecommunication network and the Internet [9]. In 3G, the amalgamation of telecommunication network, broadcasting television net and internet changes the regulation to "Integrated management" of three networks [10]. These three nets are included in the framework of uniformed mode.

It is possible to have a smooth operation process and to satisfy the demands of users. It is also possible to serve the users in the process of using the resource of digital library.

IV. ANALYSIS OF IMPLEMENTATION SCHEME

Whether entity library or digital library, functions are to satisfy the users' demanding of knowledge. Digital library is not only meets the users' learning needs, but also could be used in any time and any place. It is proposed a requirement for PTT to capture the users' context-aware environments, to decide communication form according to the environments. PTT service is a kind of information service, it involves providing user's identity and the identity of the others [10].

Using an example "a group use digital library resources online to watch a video about expert knowledge " to illustrate how to make a smoothly communication without time-delay in the context-aware environment.

A group watches experts video online, it is a way to use digital library resources. PTT online service is a to collect and then show the background information for readers. PTT technology service uses in Digital library, is to select dynamically which members can join the PTT conversation, which is based on the background of the team members' states [10]. However, PTT service is not a face-to-face conversation, so it needs to have a potential bulding service server to mark users' environment accurately. This enables users to enjoy digital library resources fully.

The user A in this group is watching video by a smart phone, because of the business relationship, user A must

be provisional interrupt online video to enter the conference room directly. PTT function keys, however, once opened, once turned off, again to re-login. On one hand to bring trouble to the users, on the other hand online videos on the use of the number of PTT users is limited. Too many people to join the group will reduce the quality of communication, affect the exchange rate, and will reduce the possibility of users re-join the group.

Therefore, PTT server is designed to have a context-aware environment, based on the user's current state to collect environmental factors and the factors that together are packaged, to form a packet, the current service system that can quickly provide the latest state of the environment information to the server-side user, this is the PTT environment used in the calculation of context-awareness. And this purpose is to capture the environments which to serve specific users, special time, special places, and to provide appropriate services [11].

When user A transferred temporarily from the state of learning to conference, He shouldn't withdraw from PTT services, but retain to use the space resources, and so on through the end of the session and then tell the server through the environmental factors to re-enter the learning state, which reaches a learning and exchange.

In this group in the communication process, Due to to-many communications that the exchange is not blocked between experts and users, happens the time delays and inadequate communication issues, PTT services solves these emergences. PTT is a two-way voice and data communications according to demand services [11]. When the expert finishes the teaching, users and experts are talking online, it will be snatch a dialogue at the same time with user A, B, C. Therefore, PTT uses the SIP-based system architecture, in the SIP-based call control program, the server forwards the user A's request and distributes receiving rights from all members' requirements to all team members. The user A presses the button and sends the call request to the server [10]. The server lookups the user B's information and forward the request to the user B. After the user B response the request to the server and the server forward the response to the user B, the user A can speak. When the user A releases the button, the call control relinquish information are transferred to the other. When the user B hope to answer the user A's calling, he will press the button and waiting for the call control allowance.

If put the PTT technology on digital library, as context-aware environment servers and SIP-based system architecture , the group users communicate on line will shorten the delaying-time, and also an opportunity will be greatly enhanced about the exchange with experts and express their views among the group members.

V. PROBLEMS AND SOLUTIONS

PTT service in the medical and other industries have already been applied, but only in the form of a simple walkie-talkie to meet the calling requirements between patients. The 3G is going to be passing into the digital library, in many respects; there are still some problems:

A. Existing Problems

*1) Charge fee.*Mobile Company set up four packages, charging according to access flow, respectively 50RMB for 500MB, 100 RMB for 2GB, 200 RMB for 5GB , 300 RMB for 10GB, and 0.01RMB per KB for the excess flow every month [12]. In other words, as to users, they have to pay at least 100RMB per month (exclusive charges on call), then are able to enjoy the digital library services through the smart phone's PTT function.

The charge of Telecom Company is a little more humanistic, inheriting the fixed broadband ADSL model which charges according to the time used. However, the kinds of charge standards are only applicable to those users who have conducted a variety of packages business provided by Tianyi , but some phones may not yet support digital libraries to meet the requirements of the PTT service.

The most low-level service provided by China Unicom 3G is 186 RMB per month, the most high-level is 1686 RMB per month, which means that a monthly income of more than 6200RMB may be willing to accept these services only. However, the most extensive users of smart mobile phones , which devote to digital library of PTT services, will be college students, white-collar workers, researchers, technical staff. For ordinary users, in particular, in terms of the majority of college students, graduate students, they all have certain difficulties in charge if they completely discard notebook to use 3G smart phones on the Internet.

(2) *Requirements of Smart Mobile Phone Terminal.* PTT service requires mobile phones in the terminal to give certain support, stakes a claim to smart phone with a complete operating system, such as Symbian, PocketPC, Palm, Linux, iPhone and other operating systems. Possibly these operating systems not the same powerful as Microsoft Windows systems, but only installing software, you can play online video resources, identify a variety of electronic document formats, and the complete literature search, data download and upload, on-line exchange operation, fully able to meet the handheld library needs of all aspects of the terminal. And smart phones have to meet the requirements of operating systems not only in hardware but also in software , then be able to support PTT operating in terms of quality and functionality, which result in the cost of smart phones production more than two or three thousands .Only the configuration of hardware and software has reached requirements, can digital library resources provider organizations need not to create module for mobile-specific data formats and associated operation , only need to add mobile access interface to the existing database, can the construction of digital library be come true.

(3)*Idleness of Resource*. In the transition state of the users' environment, there is a problem that the resources are still occupied even if users don't apply the resource temporarily, resulting in another groups not being able to enter in because of the limitations of the number of users, leading to some resources being idle. Such looks like a computer dealing with issues in thread. In the computer, the program consists of multiple executive threads, which

are a series of related instruction. In the smart mobile phone terminal, most programs contain only a single thread. The original operating system can only run one such procedure every time. Because the system can not handle two tasks at the same time, the next task must wait to be processed until the process of the first task ended. Since the latter innovation of operating system and the introduction of multi-tasking, then operating system is able to hang a program in order to run another program. By using this way to quickly switch programs, the system can run multiple programs simultaneously seemingly. However, in reality, processor has been running only a single thread, so that there will be a certain amount of resource idleness in the allocation of resources.

B. Solutions

(1) *Solutions to the existing tariff problems.* Firstly, all enterprises of the communication industry should have the sense of competition with each other. And they should make the production costs lower. As a result, the price of products turns lower, and the industry can attract more customers. Especially, for the groups of customers, who need to use the PTT service to inquiry the resources in Digital Library, specific consumer packages can be developed. And by developing the affordable consumer packages, the tariff problems can be solved. Secondly, 3G License.3G License is an operating license of the new generation of mobile communication system, which is the combination of wireless communications and the Internet and other multimedia communications. One of the aims of developing 3G technology standard is to significantly increase the communication bandwidth, which is more than ten times of the current GPRS. In other words, the Internet services provided by the present communication squeeze the already crowded communication bandwidth, so the service fee is expensive. But the large bandwidth offered by 3G, from the view of performance-price ratio, the cost per unit of flow will be reduced.

(2) The realization of PTT technology has certain requirements on the mobile terminal, which enhanced the standards on the technology, while the price of mobile phones is relatively high. To reduce the price of mobile phone, the development departments of mobile phones should improve their technology. At the same time, under the market economy condition, competition mechanism can be introduced among mobile phone brands to develop inter-industry related standards and regulate the market price.

(3)To the problem of resources idle, there are two solutions as follow. Firstly, set a waiting-timeout. In the concurrent operation, the use of multiple threads easily leads to deadlock. The method, used to solve the multi-threading problem in database, can be adopted. Optimize the algorithm in the aspect of procedure in PTT (such as the orderly resource allocation method); set a waiting-timeout in the database server of PTT and delete directly the database process when deadlock occurs.

Secondly, use the Hyper-Threading Technology. This method is to run multiple programs simultaneously in a CPU and share resources within this CPU, in theory, can be same effective as two CPU to execute two threads at the same time. So that, the CPU need to join one more logical processing unit, and the rest which is shared by users, such as integer arithmetic units, floating-point unit, second-level cache, remains unchanged. Running on a CPU with Hyper-Threading technology, the applications written in multiple threads can access to up to 30% performance increases. More importantly, the two programs can be run simultaneously on one processor without having to switch back and forth.

VI. CONCLUSION

Based on the 3G networks, the use of PTT in the digital library can make up for the shortcomings in the present functions. And as time goes on, Telecom operator will reduce the cost of 3G, open more economic packages and increase the flow of surfing and other value-added Internet services. However, there are defects in the using at present such as space occupation, lots of resources idle, and so on. These deficiencies will be made up in future with the development of technology. In short, applying the 3G technology into the digital library will be a great deal of convenience for users, while the adoption of PTT technology will also be a great opportunity in the development of the digital library.

REFERENCES

[1] Li Mingze, Zhang Yan and Tang Linchao, “3G mobile value-added operations research” in Wireless communications,vol.1, pp. 8, January 2005.

[2] Luo Jun, Zheng Xiaolin and Peng Chenglin, “Application of medical services in 3G era” in Chinese medicine equipment, vol.6, pp.37, June 2009.

[3] Chinese Internet message center（CNNIC）.Chinese development of Internet statistics report, January 2008.http://www.cnnic.com.cn,2008-03-02

[4] Hang Jizheng. 3G technology in the Digital Library Application[J] Information Exploration, vol.4, pp. 91-92, April 2009.

[5] Xu Mingjie and Li Yi, “‘PTT’ form another wireless mobile storm” in The new technology and new business ,pp 72-74, February 2005.

[6] Rui Santos Cruz, Mario Serafim Nunes, Guido Varatojo and Luıs Reis, “Push-To-Talk in IMS Mobile Environment ” 2009 Fifth International Conference on Networking and Services, pp.390.

[7] Cheng Zhiyu, “3G systems multi-user detection technology research” in Modern commercial industry, vol. 12, pp.265, 2009.

[8] Tian Yongjun and Jia GuoQin, “A GPS technology in the research and application of in the smart phones” in Computer and Network, pp.160.

[9] Li Su, “From 2G to 3G, regulators must step a big step forward” in Policy and Law，vol.9 , pp.34-36, 2007.

[10] Jenq-Muh Hsu,Wei-Bin Lain and Jui-Chih Liang, “ A Context-Aware Push-to-Talk Service ” 2008 International Conference on Multimedia and Ubiquitous Engineering , pp.586-588.

[11] Cao Peng and Yang Xuejun , “ Performance Analysis of SIP-based Push-to-Talk Service for GPRS/cdma2000 Network ,” pp.1-3.

[12] http://tech.163.com/mobile/special/00113CT3/ 517zifeihuizong.html.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 414-417

Studies on Fuzzy Comprehensive Evaluation of Trust Information System

Ping Teng, and Ping He
Department of Information , Liaoning Police Academy , Dalian, 116036 China
e-mail: Tengping-2000@163.com

Abstract — **Evaluation of an information system's success and user satisfaction are important issues in information systems research, especially for emerging online service systems on the Internet. The purpose of this study is to develop an evaluation model for question investigation from user's perspective. We have established the theoretical foundation and conceptualization of the constructs for user satisfaction with question investigation. This paper is posing the problem of how we should evaluation the reputation of information system, taking the practical conditions in China's enterprise information system (CEIS). It's also discussing the domain of information system reputation, the method for establishing dependent functions for the evaluation of information system. And the author has founded a direct method for the recognition of many factors, which is valuable in practical application.**

***Index Terms*—Information system; membership functions for reputation; comprehensive evaluation; relative membership model**

I. INTRODUCTION

Evaluation of an information system's success and user satisfaction are important research issues in the field of information management, especially for online service systems on the Internet. Evaluation models are used to understand user's needs and identify important dimensions and factors in the development of information systems in order to broaden their acceptance. With the rapid growth of the Internet and database technologies in recent years, the evaluation of Trust Infoematiob Systems (TIS) have emerged as important applications [1-5]. Hence, they have received a great deal of attention from information systems researchers, particularly those in the information retrieval and natural language processing communities.

TIS is an abstract from the practical conditions, and the comprehensive evaluation of the user's of information systems. It has become a skill, which must be grasped by the organizations, to employ subsequently qualitative and quantitative analysis to make a plan or decision according to the variations of its reputation, and to determine its qualities and the inside relationship between its trusted degree. Quantitative analysis is to discuss the inner commotion between the trusted degree of information systems and the qualities with mathematic methods, on the basis of qualitative analysis.

The purpose of this paper is to develop an effective fuzzy evaluation model for TIS based on research result of literature [1-8]. Based on the user's investigation in the degree of satisfaction to information system, we propose a practical method for the design of TIS to make appraisal user satisfaction and acceptance. As most evaluation method focus on humanity-centered evaluation, practical-centered evaluation and security-centered evaluation has attracted little attention [.5, 4]. However, if we are to build a practical RIS, we must achieve a performance level that satisfies the majority of users. Therefore, in this paper, we propose an integrated fuzzy comprehensive evaluation method of successful TIS from the user's perspective. Our goal is to analysis two questions. (1) How do individual users evaluate the success of TIS? (2) What factors influence an individual user's evaluation of a TIS success?

This paper is organized as follows: Section 2 provides a brief overview of the evaluation on TIS, setting the background for the present research, and shows the architecture of TIS in uncertainty category. Section 3 introduces the method of fuzzy comprehensive evaluation based on uncertainty category. Finally, Section 4 draws a conclusion and further work.

II. FUZZY COMPREHENSIVE EVALUATION OF TIS

A. The domain of Information Systems Reputation

Discussing domain refers to the scope and space of the studied object. As a proposition to be discussed, information systems may be considered as a discussing domain. Discussing the domain of TIS is made up of elements. The elements in the discussing domain of TIS are including the humanity of information systems (HIS), the practical of information systems (PIS), the technology service of information systems (TSIS) and the price of information system (software price) (SP). And, they are defined the quality of information systems (See figure 1).

The quality of information systems is the objective request of the quality or the product suitability, reliability, an economy of the information systems products.

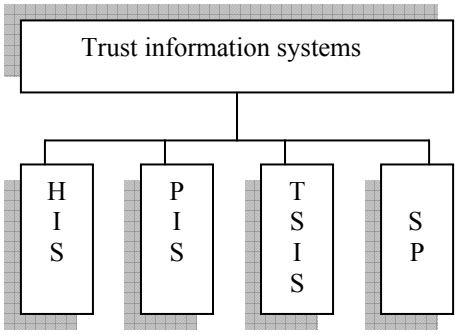


Figure 1 The structure of trust information systems

AP-PROC-CS-09CN005

It's made up of two aspects: one is economic of product, and the other is technological of products. Information systems product's price is the money representation of information systems value. TIS can be denoted by:

U = {humanity, practical, security, service and price} (1)

They are the elements in the discussing domain of organizational reputation, each of which has area boundary in its own sense. The relationship between them and the TIS is fuzzy mathematically. TIS may be described in appraisal language "very satisfactory", "both satisfactory and dissatisfactory", "dissatisfactory", which the appraisal language as the elements, an appraisal aggregation of an organization reputation is formed.

V = {very satisfactory, both satisfactory and dissatisfactory, satisfactory, dissatisfactory} (2)

There is no clear limit between the elements of appraisal aggregation, and it is only a recognition that is resulted from common practice in the process of human thinking. The process and contents of human thinking is always represented in a certain way, which may be expressed quantitatively with modern algebra to an appropriate mathematic model. To appraise the elements information systems reputation domain with approbation, the degree of subordinate appraisal aggregation of product quality, service and product price must be determined. The membership function $\mu(u)$ when information systems quality is "very satisfactory", membership function μ_1(Very satisfactory, u_1) =1, when information systems quality is "dissatisfactory", membership functions μ_1(Dissatisfactory, u_1) =1.

Membership function represents the degree of the users' appraisal of information systems quality, being any of the values between 0, 1. Humanity of information systems is subject to appraisal aggregation, the horizontal quantity formed by different membership function value is called the horizontal quantity of information systems, denoted as:

R_1= {μ_1 (very satisfactory, u_1), μ_1 (satisfactory, u_1), μ_1 (both satisfactory and dissatisfactory, u_1), μ_1 (dissatisfactory, u_1)} (3)

So it is with the horizontal quantity of membership function of PIS, denoted by expression (4) that of SIS by (5) that of PP by (6)

R_2= {μ_2 (very satisfactory, u_2), μ_2 (satisfactory, u_2), μ_2 (both satisfactory and dissatisfactory,u_2), μ_2 (dissatisfactory,u_2)} (4)

R_3= {μ_3 (very satisfactory, u_3), μ_3(satisfactory,u_3), μ_3 (both satisfactory and dissatisfactory, u_3), μ_3 (dissatisfactory, u_3)} (5)

R_4= {μ_4 (very satisfactory, u_4),μ_4(satisfactory, u_4), μ_4 (both satisfactory and dissatisfactory, u_4), μ_4 (dissatisfactory, u_4)} (6)

The determination of membership functions random appraisal of the information systems quality, the humanity of information system and the practical of software is not feasible. We may engage in study of the appraisal in the aspects of the subjective (internal factors) and the objective (user's reflection).

B. The License of Information Systems

In the design of the information systems, an organization often considers quality as the target. It is first to design the information systems according to the request of the users and the specific standard of that of the state. According to the designed quality demand, the software of information system should be made both strict and homogenized. Influenced by the stability of the technological process, the homogenizing degree of the software quality is usually reflected by the grade ratio, which is divided into the ratios of the top quality, the first-rate, the qualified software and unqualified software. The high quality ratio is the percentage of the user's satisfactory degree. So it is with other grades. How should we determine the membership function with clear logic? There is no rule for it at present. We but have to give the dependent functional mathematic model. If an organization has not got the design license of information system, then we have:

μ_1 (dissatisfactory/license, u_1) =1

μ_1 (very satisfactory/license, u_1)

= μ_1 (satisfactory/license, u_1)

=μ_1 (both satisfactory and dissatisfactory, u_1)

=μ_1 (dissatisfactory, u_1) =0

If got it, μ_1 (satisfactory/license, u_1) =1 If the higher office has not delivered it, we may regard the information system as having got it. When the ratio of high quality and the one of first –rate equals to u_1, the function of HIS is denoted by formula or expression (7)

μ_1 (satisfactory, u_1) = μ_1 (satisfactory / license, u_1)

$$\bullet \begin{cases} 0 & u \le 0.6 \\ \left[1+\left(\dfrac{u\times 100-55}{7.5}\right)^{-2}\right]^{-1.25} & u > 0.6 \end{cases} \quad (7)$$

Then

μ_1 (very satisfactory, u_1) =μ^2_1 (satisfactory, u_1)

μ_1 (dissatisfactory, u_1)

=1-μ_1 (satisfactory, u_1) (8)

μ_1 (both satisfactory and dissatisfactory, u_1)

= μ_1 (dissatisfactory, u_1) $\wedge$ μ_1 (satisfactory, u_1)

If an information system u_1=0.7, then put u_1 into formula (7), and calculating (8), we may get the horizontal quantity of the quality membership function, R_1= [0.57, 0.76, 0.24, 0.24], which can be normalized as

$$R_1 = [0.32, 0.42, 0.13, 0.13] . \quad (9)$$

C. The Means of PIS

The information system with commodity so as to serve the information management construction, and the main economic norm, which represents to provide the information system with Humanity Service (HS), is the design process. If an information system cannot provide the software with commodity, you can't begin to talk about information system service. The information system design is the essential condition that determines the recognition level of service object, to guarantee the interests of the user. The information system lowest risks have the veto power on the design of software. When the users' satisfactory degree are equal to or smaller than the lowest level:

μ_2 (dissatisfactory, u_2) = 1,

μ_2 (very satisfactory, u_2) =μ_2 (satisfactory, u_2)

= μ_2 (both satisfactory and dissatisfactory,u_2)=0.

μ_2 (satisfactory, u_2)

$$= \begin{cases} 0 & u_2 \le s_{\min} \\ \left[1 + \left(\dfrac{\dfrac{u_2 - s_{\min}}{u_2} \times 100}{65} \right)^{-2} \right]^{-1.25} & u_2 > s_{\min} \end{cases} \quad (10)$$

Only when the satisfactory degrees are larger than the lowest level, can the information system get high appraisal. The membership functional mathematic model of product service is expressed in formula (10)

In formula (10), S_{min} representing the lowest level, u_2 representing the level in the period when the information system reputation is discussed.

If the information system's lowest PIS are 0.6 them into formula (10) and with the calculation method in (8), we may get the solution of the membership functional horizontal quantity of the product service, R_2= [0.7, 0.84, 0.16, 0.26] which may be moralized as

$$R_2 = [0.37, 0.45, 0.09, 0.09] \quad (11)$$

D. Technological Service of Information System

All the services except for the product service are belonging to technological service, including "three guarantees" propaganda of product, market forecasting, and so on. In practice you may find out that the "three guarantees" of product, the ratios of carrying out contracts, and market forecasting are playing important roles in technological service. The "three guarantees" ratio of contracts carried out are the major outside technological service to the outside (the organization's consumers); and the market forecast is the major inside service (to the sections in the organization).Given u_3 stands for the ratio of three guarantees realizing; the ratio of carrying out the contracts (the ratio of contracts completing); n/N for the ratio of the accuracy of the forecasting; N for the total number of the market forecasting. The dependent functional mathematic model is shown in formula (12).

μ_3 (satisfactory, u_3)

$$= \begin{cases} 0 & u_3 \le 0.5 \\ \dfrac{n}{N} \bullet \left[\theta + \left(\dfrac{u_3 \times 100 - 50}{6} \right)^{-2} \right]^{-1.25} & u_3 > 0.5 \end{cases} \quad (12)$$

If an organization has 19 items of "three guarantees", which the users satisfied with 17 of item in a cycle, then u_3 =0.89, the ratio of contracts completing is 100%; and in this cycle, if we make a forecast for 9 kinds of products, with 7 serving as guide to planning and decision-making then the accuracy n/N =0.78. Put it into formula (12) and with the calculating and method (8), and the solution to dependent functional horizontal quantity R_3 = [0.56, 0.75, 0.25, 0.25], normalized as

$$R_3 = [0.31, 0.41, 0.14, 0.14] \quad (13)$$

E. The Prices of Software Product

According to the demand of the reformation of the economic system in our country, an organization has the right to determine the prices within a certain scope set by the state or according supply and demand in the market. Any organizations have the critical cost in the profit and loss with q_0 denoting it .In order to increase the profit, an organization tries hard to deduce critical cost, forming the sale cost of the product, denoted by u_4. Only when $q_0 - u_4$ has a bigger value, can the prices of industrial products be competitive in the marker, can the organization have higher reputation. Then the membership functional mathematic model for product prices is demoted by formula (14)

μ_4 (satisfactory, u_4)

$$= \begin{cases} 0 & u_4 \ge q_0 \\ \left[1 + \left(\dfrac{\dfrac{q_0 - u_4}{q_0} \times 100}{15} \right)^{-2} \right]^{-1.25} & u_4 < q_0 \end{cases} \quad (14)$$

When q_0 = 19.5, u_4 = 15 and putting it in formula (14) and calculating with the method in (8), we may get the solution that the horizontal quantity of dependent functions R_4 = [0.4, 0.64, 0.26, 0.26] normalized as

$$R_4 = [0.25, 0.41, 0.17, 0.17] \quad (15)$$

From above we may get:

R_1= {μ_1 (very satisfactory, u_1), μ_1 (satisfactory, u_1),

μ_1 (both satisfactory and dissatisfactory,u_1),

μ_1 (dissatisfactory, u_1)} = (0.32, 0.42, 0.13, 0.13)

R_2= {μ_2 (very satisfactory, u_2), μ_2 (satisfactory, u_2),
μ_2 (both satisfactory and dissatisfactory, u_2),
μ_2 (dissatisfactory, u_2)} = (0.37, 0.45, 0.09, 0.09)

R_3= {μ_3 (very satisfactory, u_3), μ_3 (satisfactory, u_3),
μ_3 (both satisfactory and dissatisfactory, u_3),
μ_3 (dissatisfactory, u_3)} = (0.31, 0.41, 0.14, 0.14)

R_4= {μ_4 (very satisfactory, u_4), μ_4(satisfactory, u_4),
μ_4 (both satisfactory and dissatisfactory, u_4),
μ_4 (dissatisfactory, u_4)} = (0.25, 0.41, 0.17, 0.17)

III. CONCLUSION AND FUTURE RESEARCH

We have rigorously tested the TIS instrument and found that it provides a high degree of confidence in the reliability and validity of the scales. A comprehensive model for measuring RIS is presented. In this study, we developed fuzzy comprehensive evaluation (FSE) for measuring TIS. The four membership function of TIS is ease of use, usefulness, service quality, and content quality. To enhance user satisfaction and the success of information systems, we have developed an integrated theoretical evaluation model for such systems, based on a review and synthesis of existing IS user satisfaction and technology acceptance models. We believe the proposed evaluation model provides a framework for the design of information systems from the users' perspective and that it could help increase user acceptance of information system.

ACKNOWLEDGMENT

This work was completed with the supported of the research foundation from Ministry of Public Security (Grant No. 2009YYCXLNST023). Author would like to thank professor Kang Shuhua for producing silhouettes with the developed method.

REFERENCES

[1] Ping He, The Application of Fuzzy Mathematics in Economic and Management, Liaoning Science and Technology Publishing, 1985.

[2] Taichang Shen, The System Analysis and Management Decision, Beijing: Expectation Publishing Company in China, 1984, pp.32-44.

[3] Wenji Min, Jianming Chen, Zhongyi Zhang, The Research of Information System Evaluating Index System and Method, Railway Transaction, 2000, Vol.22(5): 37-41.

[4] S. F. Abdinnour-Helm, B. S. Chaparro, and S. M. Farmer, "Using the end-user computing satisfaction (EUCS) instrument to measure satisfaction with a Web site," Decision Sciences, vol. 36, pp. 341-364, May 2005.

[5] J. Allan, B. Carterette, and J. Lewis, "When Will InformationRetrieval Be "Good Enough"? - User Effectiveness As a Function of Retrieval Accuracy," in Proceedings of ACM SIGIR ,2005.

[6] J. E. Bailey and S. W. Pearson, "Development of a Tool for Measuring and Analyzing Computer User Satisfaction," Management Science, 1983, vol. 29, pp. 530-545.

[7] J. Baroudi and W. J. Orlikowski, "A Short-From Measure of User nformation Satisfaction: A Psychometric Evaluation and Notes on Use," Journal of Management Information Systems, Spring ,1988vol. 4, pp. 44-59.

[8] Lenat, D.B. and Guha, P.V., Building Large Knowledge Based Systems: Representation and Inference in the CYC Project, Addison Wesley, 1990.

[9] R. Milner, Communicating and mobile systems: the p--calculus, Cambridge university press, 1999.

[10] Cardelli I., Gordon A D., Mobile Ambients, in M.Nivat, editor, Foundations of Software Science and Computational Structures, LNCS No.1378, Springer Verlag, 1998, pp.140—145.

[11] He Ping, Fuzzy comprehensive Appraisal of the Reputation of China' Business Organization, Knowledge Economy Meets Science and Technology-KEST2004, 2004, pp. 473-477.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 418-422

On Memory Management of Tree-bitmap Algorithm for IP Address Lookup

Yagang Wang[1,2] , Huimin Du[2] ,and Kangping Yang[2]
[1]Computer school, Xidian University, Xian 710071, China
Email: wangyg@xupt.edu.cn
[2]Department of Computer Science, Xian Institute of Posts and Telecommunications, Xian 710121, China
Email: {fv, yangkp}@ xupt.edu.cn

***Abstract*— A Memory Management Unit (MMU) is adopted in the implementation of the Tree-bitmap algorithm for IP address lookup, and its memory allocation policy is vital to the performance of incremental update of Routing Information Base (RIB). Using the RIB database of active routers, a Tree-bitmap based IP address lookup table is constructed, and the memory allocation pattern of the incremental update scheme is analyzed. Based on the analysis, a reference free memory arrangement model is proposed. This model serves as a practical guide for the implementing the MMU for an incremental update version of the Tree-bitmap algorithm. Experimental results show that the proposed model is of good quality and its steady distribution can be applied to most RIBs in today's routers, and can be used to reduce the amount of memory copying by as much as 85%.**

***Index Terms*—tree bitmap, IP address lookup, memory management, routing information base, incremental update**

I. INTRODUCTION

With the ever increasing expansion of the Internet, a huge number of new network applications emerged. These applications in turn put forward a new performance requirement on the Internet itself. As the core device of the Internet, IP routers must meet with these performance requirements. An IP router receives IP packets from an ingress port, lookups the next hop port based on the destination IP address and the RIB (Routing Information Base), and finally forwards the packet to the proper egress port. With the ever growing size of the RIB and the introduction of IPv6, the efficiency of the address lookup algorithm has become a dominating factor in router performance improvement.

In order to meet the ever increasing performance demand of IP address lookup, many IP address lookup algorithms have been proposed in recent years. A detailed summary of these algorithms can be found in [1,2]. Among these algorithms, the Tree-bitmap algorithm is one of the most efficient. This algorithm has lower memory usage and fast incremental update performance [3-5]. In the hardware/software implementation of Tree-bitmap, all the child trie nodes of a parent node, and all the next hop information of a trie node, should be stored in contiguous memory locations. As a result, a large amount of memory movement and copy may be committed during incremental RIB update. In order to cope with the complexity of memory management in the incremental update process, a Memory Management Unit (MMU) is adopted in the implementation of the Tree-bitmap algorithm. However, too much number of memory copies and memory movements will significantly degrade the update performance.

Based on our experiments, we found that, during the process of incremental prefix updates, the amount of the memory movement has closely relationship with the initial allocation status of the free memory spaces. In another word, the initial free memory arrangement is a key issue for the update performance in Tree-bitmap algorithm, and this is just the main research subject of this paper.

In this paper, based on the analysis of the memory management algorithm of Tree-bitmap, along with the real life RIB data acquired from the routerview.org website [8], the distribution of memory allocation requirement of different block size is presented, which can be a reference free memory allocation model for the Tree-bitmap algorithm and can dramatically reduce the cost of memory movement for routing prefix update.

The rest of the paper is organized as follows. In Section 2, we review the background and related research about the Tree-bitmap algorithm. In Section 3, the memory management algorithm of Tree-bitmap is discussed. In Section 4, a new reference model is proposed based on experimental results acquired from real RIB archives. Finally, a conclusion is reached and the further study is presented as well.

II. RELATED WORK

A. Tree-bitmap Algorithm

Eatherton proposed a data structure for longest prefix lookup based on multi-bit expanded tries [3,4]. This structure is scalable in terms of table size, lookup speed, update speed, and flexibility to adapt to the next generations of memory technology. There are four key ideas in the Tree-bitmap algorithm. 1) All the child trie nodes of one parent node should be stored contiguously. There is only one Child Node Pointer in the parent node, and all the child trie nodes can be referenced by this

Manuscript received October 26, 2009.
This work was partially supported by a grant from the National Science Foundation of China (No. 60976020).
Corresponding author: Yagang Wang, Email:wangyg@xupt.edu.cn.

Child Node Pointer and an address offset. 2) There are two bitmaps in a trie node, the Internal Prefix Bitmap and Extending Path Bitmap. An Internal Prefix Bitmap is used to identify which prefix in the trie node is valid by set the corresponding bit to "1", and the Extending Path Bitmap is used to identify the valid extending child trie nodes. 3) The trie nodes should be kept as small as possible to reduce the required memory access times for a given stride. 4) All the next hop information of the prefixes related to one trie node should also be kept in contiguous memory location, with the Next Hop Pointer pointing to the first item of the next hop information array.

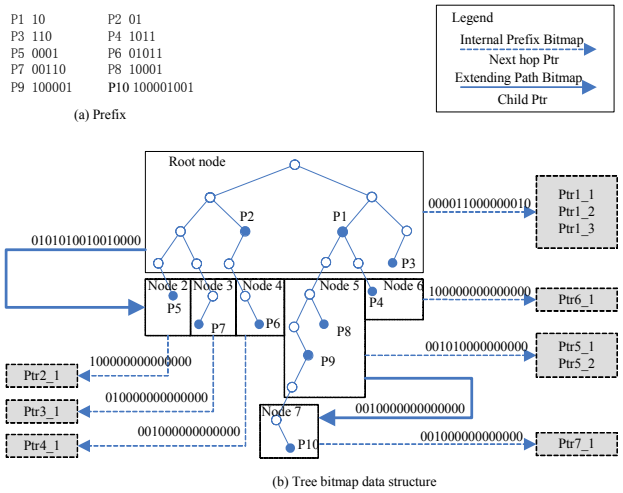


Figure 1. Data structure for Tree-bitmap algorithm. There are four fields in a trie node: Internal Prefix Bitmap, Next Hop Pointer and Extending Path Bitmap, Child Node Pointer.

For example, as shown in Fig.1, (a) is the prefix database, and (b) is the corresponding Tree-bitmap data structure with a stride of 4. There are four fields in each trie node: the Extending Path Bitmap, the Child Node Pointer, the Internal Prefix Bitmap and the Next Hop Pointer. In the Root node, the five "1"s in the Extending Path Bitmap 0101010010010000 identify the paths from the Root node to $Node_2$, $Node_3$, $Node_4$, $Node_5$ and $Node_6$ respectively. These five child trie nodes are stored in contiguous memory locations, with the Child Node Pointer in Root node pointing to their base address, say $Node_2$. The three "1"s in Internal Prefix Bitmap 000011000000010 of the Root node identify the prefixes P_1, P_2 and P_3 associated with this node. The next hop information Ptr1_1, Ptr1_2, Ptr1_3 of P_1, P_2 and P_3, are stored in contiguous location with the Next Hop Pointer pointing to their base address, say Ptr1_1.

As can be seen from the above example, the storage of the child trie nodes and next hop information in contiguous locations helps to significantly reduce the storage cost. However, the corresponding memory management scheme suffers from high complexity during incremental update to keep the character of contiguous storage. In the worse case, a prefix update may cause 1156 memory accesses as described in [3].

B. Memory Management of Tree-bitmap Algorithm

A memory management scheme is introduced to manage the memory allocation in the reference design [3]. To make incremental update more effective, the variable length memory allocation method is adopted, as illustrated in Fig.2. For the reference design in [3], there are 17 different possible allocation block sizes. The minimum allocation block is 2 nodes which includes one allocation block header. The maximum allocation block size is 18 nodes. This occurs when the child array contains 16 nodes, the required allocation header node, and an internal node for the parent node. There are 17 allocated memory spaces Z_N ($1<=N<=17$) for memory allocation block size of N+1. For each allocated memory space Z_N, there are two end pointers α_N, β_N, pointing to the start address and end address of Z_N. When a memory block with size N+1 is to be allocated, two end pointers, α_N and β_N should be modified to reflect the change of memory space Z_N. Allocated Memory spaces for different block size should never inter-mingle. Therefore, there are free spaces, namely as F_1, F_2, ... , F_{17}, exist between the adjacent allocated memory spaces.

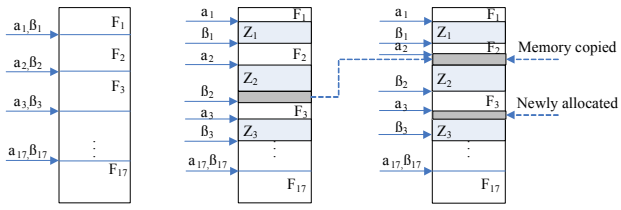


Figure 2. Memory management of Tree-bitmap algorithm for IP address lookup.

When an allocation block of N+1 is needed, 1) if there is room to extend the size of allocated memory space for that block size either up or down, then allocate the new node in the adjacent free space F_N or F_{N+1}, and adjust the end pointer α_N or β_N accordingly. 2) If the memory space can not be extended directly, a linear search is done in both directions until a free space F_M is founded large enough for the desired allocation size. Then, for whatever distance from the free space F_M, to the memory space Z_N, the desired free space have to be passed by copying the necessary number of allocation block in each allocated memory space from one end to the other. At the same time, for each allocated block moved, the parent of the block must have its child address pointer adjusted to reflect the new location.

Consider the example in Fig.2, suppose a block containing 4 nodes is to be allocated, and the nearest free space exists between α2 and β1, say F_2 (we will assume the free space F_2 is at least 6 nodes in size). To allocate memory, two blocks from the 3 node memory space (Z_2) will be moved from the bottom of the 3 node memory space to the top of that space. Then α_2 and β_2 will be shifted up 6 memory locations and α_3 will be shifted up 4 memory locations to make room for the newly allocated block.

It is usually simple to free a memory block of size N+1. For the majority of the cases, the deallocated block will

be inside the memory space Z_N and not on the edge. For this typical case, simply copy one of the blocks at the edge of Z_N into the free block. Then adjust the appropriate end pointer. If the deallocated block is on the edge of Z_N, then all that needs to be done is to adjust the end pointer.

III. Free Memory Arrangement Model for Tree Bitmap Algorithm

A. Main Problems

In the implementation of tree-bitmap algorithm, the main cost of incremental update results from the memory allocation method, which will cause a large number of the memory movement in the worst case. For example, in the worst case, an 18 nodes block is required while Z_{17} runs out of adjacent free space in F_{17}, and all the free space available is at the other end of memory space, say F_1. Then a certain number of free spaces must be passed from end to end for each allocated memory space from Z_1 to Z_{16}. In [3], the total memory access number is calculated as 34*7*2. But in fact, it needs more memory movement than that is presented in [3]. The reason lies in the fact that for each allocated memory space Z_N, the memory movement size must be a multiple of (N+1). The memory movement size in each allocated memory space is listed in Table I, so the total memory access number is 1367*2=2734, which is larger than the original number

TABLE I.
MEMORY MOVEMENT IN EACH ALLOCATED MEMORY SPACE Z_N

N	1	2	3	4	5	6	7	8	9
amount	108	108	108	105	102	98	96	90	90
N	10	11	12	13	14	15	16	17	Total
amount	88	84	78	70	60	48	34	-	1367

presented in [3] for the worst case scenario.

From the analysis of the worst case scenario, we can see that, the worst case results from the imbalance distribution of the free memory spaces. With the poor arrangement of the free spaces in advance, the distribution of the free spaces becomes more and more imbalance for incremental RIB updates. It is just the imbalance distribution of free space that causes the large amount of memory movements in RIB update. Based on this idea, this paper mainly deals with the following issues:

1) The comparisons of the memory movement amount for different initial free memory space arrangement;

2) The introduction to a practical free memory arrangement model, which can significantly reduce the total memory movement for RIB update.

B. Free Memory Arrangement Policies

Based on the previous discussion, it is clear that the initial free memory arrangement is a key issue to the update performance in Tree-bitmap algorithm. In this section, five different free memory arrangement policies and a performance comparison are presented.

In order to clarify the discussion below, the following assumptions are adopted.

1) A Next Hop information item and a trie node are same in size;

2) Z_N and F_N ($1<=N<=16$) are used to indicate 16 allocated memory spaces and 16 free memory spaces respectively. Z_N is used to store the memory block N+1, while Z_N and F_N is adjacent.

3) There are three main prefix update operations: insertion, deletion and modification. Due to the fact that the deletion and modification of a prefix update will not cause much memory movement, only insertion of a prefix in RIB is considered to evaluate the memory movement cost.

Five simple free memory arrangement policies are as following:

1) All the free spaces are equal in size.

2) All the free memory is allocated to F_1, that is, F_2, F_3… F_{16} are zero in size.

3) All the free memory is allocated to F_{16}, that is, F_1, F_2… F_{15} are zero in size.

4) All the free memory is allocated to F_8, that is, F_1, F_2… F_7, and F_9, F_{10}…F_{16} are zero in size.

5) All the free spaces are allocated to F_1, F_2…F_{16}, with their size in coincide with a certain memory requirement distribution for block size 2, 3…17. For example, a reference memory requirement distribution is illustrated in Table II.

TABLE II.
MEMORY REQUIREMENT DISTRIBUTION FOR DIFFERENT BLOCK SIZE N

N	2	3	4	5	6	7	8	9
percent	68	10	5	4	2	1	1	1
N	10	11	12	13	14	15	16	17
percent	1	1	1	1	1	1	1	1

Using above five policies, a series experiment have been done to check the real performance for different free memory arrangement policies. In our experiments, the memory can hold 128K trie nodes in all. With the same RIB update data, when the memory is under 100% utilization, the total amount of memory movement for each policy is summed up, and listed in Table III.

TABLE III.
MEMORY MOVEMENT FOR FIVE POLICIES

Policy	#Memory Request	#Memory Movement (Block)	Performance (Normalized)
1	39218	435945	1.00
2	39216	828425	1.90
3	39227	1310740	3.01
4	39228	676201	1.55
5	39218	65267	0.15

It is clear that when the size of free space F_N is in agree with the corresponding memory request distribution of the block size (N+1), the total memory movement overhead is rather lower than the uniform distribution of

the free space. In terms of the total memory movement amount of each policy, the final column in table III shows the normalized performance with policy 1 (uniform distribution) as a baseline.

C. Reference Model for Free Memory Arrangement

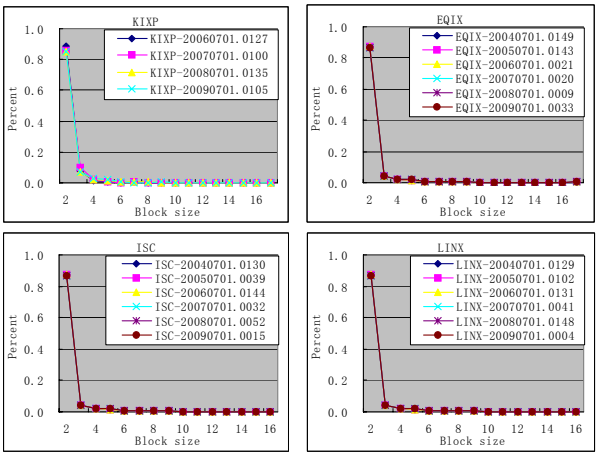
It is clear that, the fully utilization of the memory with little memory movement can be achieve by rational free memory arrangement according to a specific memory requirement distribution for different block size. The development of such a specific memory requirement distribution is then become the key issue to be discussed.

There are three main steps to form this memory requirement distribution:

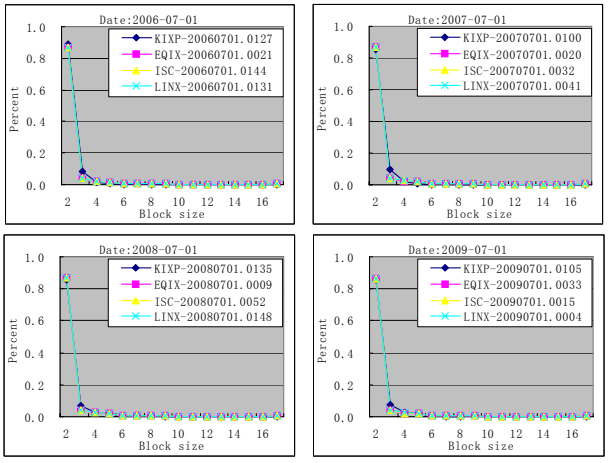
1) Obtain the active RIB data, for example, from the Route Views Archive Project Page [6].

2) Construct the Tree-bitmap data structure.

3) Sum up the number of the occurrence of different size of memory block respectively.



(a) Memory requirement distribution for RIB of a BGPds in different times.



(b) Memory requirement distribution for RIB of different BGPds in the same time.

Figure 3. Memory requirement distribution of different block size. The distribution of different block size remains steady from year 2004 to 2009, for different zebra bgpd routers.

In this paper, Four RIB archives of Equinix Ashburn, ISC (PAIX), KIXP and LINX zebra BGPds on RouterView are selected as the experimental data source. We choose the first RIB data of July, 1st each year from year 2004 to 2009 (for KIXP, data available only from 2006 to 2009).

To calculate the number of different block size, the Internal Prefix Bitmap and Extending Path Bitmap is used. By counting the occurrence of "1" in the bitmap, we can determine the size of the desired memory block, in which to hold the child tree nodes or the next hop information. For example, an Internal Prefix Bitmap 101010000000000 has three "1"s in itself, so the desired memory block which contains an extra memory management header and three next hop information items, should be allocated in Z_3. This will lead to a memory requirement of block size 4. As for an Extending Path Bitmap 1000110010100001, it has six "1"s, so a memory block which contains an extra memory management header and six tree nodes, should be allocated in Z_6. This will result in a memory requirement of block size 7. By computing all the bitmap in the Tree-bitmap data structure, the total number of each memory block in different size can be obtained respectively.

As a result, all these statistical data can be used to form a reference distribution of memory requirement for different block size.

In Fig.3, the experimental results of memory requirement distribution of different block size are grouped into two series:

a) Memory requirement distribution for RIB of one BGPd in different times.

b) Memory requirement distribution for RIB of different BGPds in the same time.

As shown in Fig.3, based on the construction of Tree-bitmap data structure from the real RIB data, we obtain a memory requirement distribution for active routers in different periods. As we can see, the block of size 2 is dominated for about 87 percent, and then block of size 3 for about 5 percent, the others for about 8 percent in total.

Furthermore, a more important discovery is that the distribution model remains fairly steady from year 2004 to year 2009 and for different zebra BGPd routers as well.

As a result, the block size distribution illustrated in Fig.3 can also served as a reference memory requirement distribution for the free memory arrangement model. A free memory arrangement model is proposed in Table IV, which is an average distribution acquired from Fig.3.

TABLE IV.
FREE MEMORY DISTRIBUTION FOR DIFFERENT BLOCK SIZES

N	2	3	4	5	6	7	8	9
percent	86.8	5.1	2.3	1.6	0.9	0.7	0.5	0.5
N	10	11	12	13	14	15	16	17
percent	0.2	0.2	0.2	0.2	0.1	0.2	0.1	0.4

IV. CONCLUSION AND FURTHER STUDY

In the implementation of Tree-bitmap algorithm, the initial free memory arrangement is a key issue to the prefix update performance. Based on the active RIB data from the Route Views Archive Project Page, a reference free memory distribution is obtained from a large amount

of statistical data, which remains fairly steady from year 2004 to year 2009 and for different routing information bases. By using this free memory arrangement mode, the memory copy amount dramatically reduced by 85 percent compared with the average allocation method.

The future study will focus on the development of a more precious model, with optimizations for Tree-bitmap algorithm are concerned.

ACKNOWLEDGMENT

The authors are grateful to the members of the High Performance Route Group of Xian Institute of Posts and Telecommunications for their efforts to this paper.

REFERENCES

[1] M Sanchez, E W Biersack, W Dabbous. Survey and taxonomy of IP address lookup algorithms [J]. IEEE Network, 2001, 15(2): 8–23

[2] Marcel Waldvogel, George Varghese, Jon Turner, and Bernhard Plattner. Scalable high-speed prefix matching[J]. Transaction on Computer Systems, 19(4):440–482, November 2001.

[3] W.N.Eatherton, Hardware-based Internet Protocol Prefix Lookups [D]. master thesis, Washington University in St. Louis, 1998.

[4] W. Eatherton, Z. Dittia, and G. Varghese, Tree bitmap: Hardware/software ip lookups with incremental updates[J]. in ACMSIGCOMM Computer Communications Review, 2004, 34(2)

[5] David E Taylor, Jonathan S Turner, John W Lockwood, et al. Scalable IP Lookup for Internet Routers [J]. IEEE Journal on Selected Areas in Communications, 2003, 21:522-534.

[6] Route Views Archive Project Page.[EB/OL]. [2009-08-07]. http://archive.routeviews. org/

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 423-426

Design and Optimization of Cluster Supply Chain Based on Genetic Algorithm

Chunling Liu[1], Jingyi Chen[2], and Aping Yuan[2]
[1]School of Electronics and Information, Wuhan University of Science and Engineering, Wuhan, China
Email: liuchunring@yahoo.com
[2]School of Economics and Management, Wuhan University of Science and Engineering, Wuhan, China

Abstract—**Recent researches regarding supply chain design mainly focus on a limited tier in single supply chain, which only take into account vertical cooperation and ignore the across-chain horizontal one. This paper, based on cluster supply chain, provides a novel framework and approach to design cluster supply chain without across-chain horizontal cooperation, then by introducing item allocation proportion of vertical and horizontal cooperation, the cluster supply chain design with across-chain horizontal cooperation is developed, then presents a hybrid method to find solution, at last, computational study is presented to investigate values of decision variables and their influence on cluster supply chain design.**

Index Terms—**supply chain management, Genetic algorithm, inventory management, across-chain coordination**

I. INTRODUCTION

Intelligent ubiquitous IT policy and its industries services are attracted more attention with the need for increased agility and flexibility in the manufacturing industry (Fletcher et al, 2002). Therefore, some specific organizations, such as Four Party Logistics (4PL), emerge and offer firms relevant services for their quick response to ever-changing market. In the real business world, the relationship among firms becomes more complex and uncertainty, the 4PL are playing and will play an important role in providing this kind of intelligent ubiquitous business model design, because the rule of competition between one firm Verse another is replaced by a chain verse another chain (Christopher, 2005), the cooperation is the same, where does it occur? With the further development of industrial and specialization division, industrial cluster provide an environment to makeup multi-chains and promote their member cooperation between them in order to implement leagilty strategy for sharpening their edge of competitive advantages (Kaufman & Rousseeuw, 1990; Punj & Stewart, 1983). Moreover, ever-changing market demand also forces firms to adopt coordination policy from firm-wide cooperation to chain-wide cooperation, and even to across-chain cooperation so that firms can survive and thrive. On this basis, we refer to multiple of single supply chains located in industrial cluster as cluster supply chain. Design of cluster supply chain with across-chain horizontal cooperation, in this paper, refers to more than one focal enterprises not only design their own individual single chains, but design the interlinked parts (i.e. across-chain horizontal cooperative components) of the two single chains as well. Therefore, this paper focuses on how to design intelligent model of the cluster supply chain of this kind for two core firms by 4PL.

In this paper, based on cluster supply chain, we provide a novel framework and approach to design cluster supply chain with across-chain horizontal cooperation. The remaining parts of the paper are organized as follows. Section "Literature review" give a brief explanation of cluster supply chain. The cluster supply chain design problem is formulated and discussed in Section "Problem presentation". Comprehensive explanation of the proposal GA approach is given in Section "Model algorithm" followed by discussion of computational experiments in Section "Illustration examples". Finally, concluding remarks are outlined in section "Conclusion and future research".

II. PROBLEM PRESENTATION

Considering cluster supply chain with two single chains, assume that each individual single chain consists of one supplier, one manufacturer and one retailer, and produce the similar or the same products among the two single chains. Meanwhile we do not take into account that the two single supply chains exist direct competition at the echelon of supplier, manufacturer and retailer. Therefore, cluster supply chains design need solving the several problem: 1) for certain product, if there is an across-chain replenishment relationship between supplier at one chain and manufacturer at the other chain, or manufacturer at one chain and retailer at the other chain; 2) determine supplier's item type and batch; 3) determine transportation routing and transportation batch between suppliers and manufacturers; 4) determine manufacturer's production batch; 5) determine transportation routing and transportation batch between manufacturers and retailers.

III. MODEL FORMULATION

The following notation is used in the formulation of the model.

j Index set of single supply chains, $j \in \{1, 2, \}$

i Index set of products available to manufacturer at single supply chains, $i \in \{1, 2, \cdots, I\}$.

$direct\_c_{ji}$ Transportation cost of delivery per unit of product i from supplier to manufacturer at single supply

Footnotes: surported by CIMS/863 Project: 2009AA04Z152, China Post-doctoral Special Project (200801312)), The Educational Department Project of Hubei Province (Z20081702, B200717001)

chain *j*.

c_j Maximum transportation capacity level for supplier shipping product *i* to manufacturer at single supply chain *j*.

$direct\_f_{ji}$ Fixed cost of per unit of product *i* for opening and operating among supplier and manufacturer at single supply chain *j*.

m_{ji} Manufacturing cost of per unit of product *i* at single supply chain *j*.

cap_{ji} Consumption production capacity for manufacturer producing one unit of product *i* at single supply chain *j*.

M_j Maximum production capacity level of manufacturer at single supply chain *j*.

$direct\_p_{ji}$ Transportation cost of delivery per unit of product *i* from manufacturer to retailer at single supply chain *j*.

p_j Maximum transportation capacity level of manufacturer at single supply chain *j*.

$direct\_g_{ji}$ Fixed cost of per unit of product *i* for opening and operating among manufacturer and retailer at single supply chain *j*.

h_{ji} Warehouse capacity needed for retailer stocking one unit of product *i* at single supply chain *j*.

INV_j Maximum warehouse capacity level of retailer at single supply chain *j*.

d_i Total market demand for product *i*.

x_{ji} Transportation volume of shipping product *i* from supplier at single supply chain *j* to manufacturer at single supply chain *j* or other single supply chain.

y_{ji} Transportation volume of shipping product *i* from manufacturer at single supply chain *j* to retailer at single supply chain *j* or other single supply chain.

In comparison with the aforementioned model, model of cluster supply chain design with across-chain horizontal cooperation is that the cluster supply chains not only exists inter-chain vertical cooperation between upstream and downstream firms along each individual single supply chain, but also across-chain horizontal cooperation as well. This kind of across-chain horizontal cooperation occurred among the two single supply chains, which avoid being out of stock or overstock so as to turn to help from the extra supply channel or emergency order channel, while the regular supply and order channel via vertical co-operational supply pipeline unable do well. For the two single supply chains located at industrial cluster, so the two single chains are characterized by production similarity, which leads to two single supply chains substitute their products with each other, the model with two single chains is showed in Fig. 1.

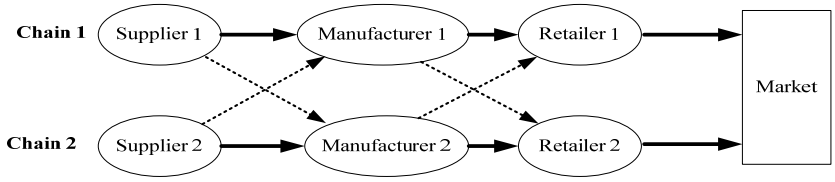


Fig. 1 Cluster supply chain system with across-chain horizontal cooperation

Let $cross\_c_{ji}$ be the transportation price of supplier at the single supply chain *j* send product *i* to manufacturer at another single supply chain by the way of across-chain shipment, and $cross\_p_{ji}$ be the production price of manufacturer at the single supply chain j produce product *i* to retailer at another single supply chain by the way of across-chain model, and $cross\_f_{ji}$ be the fixed cost of supplier at the single supply chain *j* provide product *i* to manufacturer at another single supply chain by the way of across-chain model, and $cross\_g_{ji}$ be the fixed cost of manufacturer at the single supply chain *j* provide product *i* to retailer at another single supply chain by the way of across-chain model.

α is a proportion of a firm allocating its supply items between vertical (α) and horizontal cooperation pipeline ($1-\alpha$). In addition, two new decision variables v_{ji} and w_{ji} are introduced, where v_{ji} is a binary variable of supplier at single supply chain j providing product *i* to manufacturer at the other single chain through across-chain model (v_{ji}=1 if supplier at single supply chain *j* serving product *i* to manufacturer at the other single chain through across-chain model and 0 otherwise), and w_{ji} is a binary variable of manufacturer at single supply chain *j* providing product *i* to retailer at the other single chain through across-chain model (w_{ji}=1 if manufacturer at single supply chain *j* serving product *i* to retailer at the other single chain through across-chain model and 0 otherwise), the other variables aforementioned remain intact, so the cost for the single supply chain *j* with across-chain cooperation is represented as follows.

$$\begin{aligned}\min TC_j = &\sum_i^I \alpha \cdot x_{ji} \cdot direct\_c_{ji} + \sum_i^I (1-\alpha) \cdot x_{ji} \cdot cross\_c_{ji} \\ &+ \sum_i^I m_{ji} x_{ji} + \sum_i^I \alpha \cdot y_{ji} \cdot direct\_p_{ji} \\ &+ \sum_i^I (1-\alpha) \cdot y_{ji} \cdot cross\_p_{ji} \\ &+ \sum_i^I direct\_f_{ji} \cdot u_{ji} + \sum_i^I cross\_f_{ji} \cdot v_{ji} \\ &+ \sum_i^I direct\_g_{ji} \cdot z_{ji} + \sum_i^I cross\_g_{ji} \cdot w_{ji}\end{aligned}$$

Thus the total cost of cluster supply chain design with across-chain horizontal cooperation can be formulated

$$\begin{aligned}\min TC = & \sum_{j}^{2}\sum_{i}^{I}\alpha\cdot x_{ji}\cdot direct\_c_{ji} \\ & +\sum_{j}^{2}\sum_{i}^{I}(1-\alpha)\cdot x_{ji}\cdot cross\_c_{ji}+\sum_{j}^{2}\sum_{i}^{I}m_{ji}x_{ji} \\ & +\sum_{j}^{2}\sum_{i}^{I}\alpha\cdot y_{ji}\cdot direct\_p_{ji} \\ & +\sum_{j}^{2}\sum_{i}^{I}(1-\alpha)\cdot y_{ji}\cdot cross\_p_{ji} \\ & +\sum_{j}^{2}\sum_{i}^{I}direct\_f_{ji}\cdot u_{ji}+\sum_{j}^{2}\sum_{i}^{I}cross\_f_{ji}\cdot v_{ji} \\ & +\sum_{j}^{2}\sum_{i}^{I}direct\_g_{ji}\cdot z_{ji}+\sum_{j}^{2}\sum_{i}^{I}cross\_g_{ji}\cdot w_{ji}\end{aligned} \quad (1)$$

Subject to: (2)-(5),

$$\alpha x_{1i}\cdot u_{1i}+(1-\alpha)\cdot x_{2i}\cdot v_{2i}\geq y_{1i}\ \ \forall i \quad (2)$$

$$\alpha\cdot x_{2i}\cdot u_{2i}+(1-\alpha)\cdot x_{1i}\cdot v_{1i}\geq y_{2i},\forall i \quad (3)$$

$$\sum_{j}^{2}\alpha\cdot y_{ji}\cdot z_{ji}+\sum_{j}^{2}(1-\alpha)\cdot y_{ji}\cdot w_{ji}=d_i,\forall i=1,2,...,I \quad (4)$$

$$\sum_{i}^{I}u_{ji}\geq 1\ \ \forall j=1,2 \quad (5)$$

$$\sum_{i}^{I}z_{ji}\geq 1\ ,\quad \forall j=1,2 \quad (6)$$

$$x_{ji}\geq 0\ ,\ \forall i,j \quad (7)$$

$$y_{ji}\geq 0\ ,\quad \forall i,j \quad (8)$$

$$u_{ji}\in\{0,1\}\ ,\ \forall i,j \quad (9)$$

$$v_{ji}\in\{0,1\}\ ,\ \forall i,j \quad (10)$$

$$z_{ji}\in\{0,1\}\ ,\ \forall i,j \quad (11)$$

$$w_{ji}\in\{0,1\}\ ,\ \forall i,j \quad (12)$$

The total costs (1) equally made of : the costs of shipments from supplier to manufacturer with vertical and horizontal cooperation, the costs of production of manufacturer for each individual single supply chains, the costs of shipment from manufacturer to retailer with vertical and horizontal cooperation, the fixed costs occurred between supplier and manufacturer with vertical and horizontal cooperation, the fixed costs occurred between manufacturer and retailer with vertical and horizontal cooperation. Constraint set (2) guarantees that the orders from all manufacturer should be satisfied by the supplier at the first single supply chain, and constrain set (3) ensures that orders from all manufacturer should be satisfied by the supplier at the second single supply chain. Constraint set (4) represents market demand restriction. Constraint sets (5) and (6) ensure existence of vertical and horizontal cooperation among cluster supply chain respectively. Constraint set (7-12) enforces the non-negativity and integrality restrictions on the corresponding variables.

IV. MODEL ALGORITHM

The objective for this model is the minimization of system-wide overall cost that could be broken down into fixed investment cost, variable operating cost, vertical fixed shipping cost, horizontal fixed shipping cost, vertical variable shipping cost, horizontal variable shipping cost, in-process inventory cost etc. The model with mixed non-linear program (MNILP) can be computed by Lagrange algorithm or other comprehensive algorithms (Daniel & Rajendran, 2005), but these methods are inefficient due to more variables and constraints existing in one model. For decisions are made on a set of qualitative variables, a genetic algorithm is applied to qualitative policy variables, a mixed integer programming solves the approximate model for given policy variables resulted from the genetic algorithm, and simulation is used to calculate the optimal solution of cluster supply chain.

V. ILLUSTRATION EXAMPLES

The studied case refers to one of Chinese industrial clusters where there is a cluster supply chain with two single supply chains (j=2), each single supply chain contains one supplier, one manufacturer and one retailer, the two single supply chains serve the same market with two products, the values are shown in Tab.1, through software Matlab, we computed the two above problems.

TABLE. 1(A) DATA OF PARAMETERS

Value of parameter		$direct\_c_{ji}$		$cross\_c_{ji}$		$direct\_f_{ji}$		$cross\_f_{ji}$		$direct\_p_{ji}$		$cross\_p_{ji}$		m_{ij}	
Single chain		1	2	1	2	1	2	1	2	1	2	1	2	1	2
Product	1	10.5	10.0	11.5	11.0	120	110	130	120	10.0	10.5	11.0	12.0	32	31
	2	14.0	13.5	15.0	15.0	150	145	180	170	12.0	13.0	13.5	14.0	38	37

TABLE 1(B) DATA OF PARAMETERS

Value of parameter		cap_{ji}		h_{ji}		$direct\_g_{ji}$		$cross\_g_{ji}$		d_i
Single chain		1	2	1	2	1	2	1	2	
Product	1	1.5	1.5	1.0	1.0	90	95	110	115	350
	2	3.0	3.0	3.0	2.5	140	150	160	160	380

TABLE1 (C) DATA OF PARAMETERS

Value of parameter		M_j	c_j	p_j	INV_j
Single chain	1	1100	360	360	1080
	2	1200	380	375	1150

TABLE 2 RESULT OF CLUSTER SUPPLY CHAIN DESIGN WITH AND WITHOUT ACROSS-CHAIN HORIZONTAL COOPERATION

α	Binary variables of across-chain horizontal cooperation										Connection lines (K)	Inventory		Ratio (TC/K)
	Time-consumed (min)	Target value(TC)	v				w					Product 1	Product 2	
			v_{11}	v_{12}	v_{21}	v_{22}	w_{11}	w_{12}	w_{21}	w_{22}				
0.50	2	45 337.50	1	0	0	1	1	1	1	1	12	10	0	3 778.125
0.60	2	45 174.20	0	1	1	1	1	1	1	1	14	5	0	3 226.729
0.70	2	44 785.85	0	1	1	1	1	1	1	1	14	0	0	3 198.989
0.80	2	44 605.60	0	1	1	1	1	1	1	1	14	0	0	3 186.114
0.90	3	44 364.85	1	0	0	1	1	1	1	1	12	3	0	3 697.071
1.00	1	43 112.50	0	0	0	0	0	0	0	0	6	0	0	7 185.417

As for the goal function (12), let the weighted parameter α endowed by different values ($1 \geq \alpha \geq 0$), and the step is 0.05, and compute values of decision variables and of the goal functions with different weighted parameters. It is found that for $\alpha < 1-\alpha$ (or $0.5 > \alpha > 0$), the cost of cluster supply chain design with across-chain cooperation is greatly higher than that without across-chain cooperation, and for $\alpha \geq 1-\alpha$, the situation is reverse, that is to say that the cluster supply chain design with across-chain horizontal cooperation incurred less cost. This change implies there exists close relationship between the total cost and weighted parameter α. Although the parameter α is effected by a couple of factors, such as inventory level, order quantity from downstream in regular channel, and demand quantity in horizontal emergency channel, the parameter α is playing the most important role in impacting the total cost. In other words, the allocation proportion of vertical and horizontal cooperation (α : $1-\alpha$) is linked to its own cooperation cost incurred vertically and horizontally.

In real business world, vertical coordination along cluster supply chain is a long term and orientated-strategy, while horizontal cooperation is a temporary and short term contract. Due to that strategy is overall and long term relationship, thus the operation cost is relative low, while horizontal cooperation is extra and temporary one, the operative cost higher. Although α and $1-\alpha$ is the item allocation proportion of vertical and horizontal cooperation in cluster supply chain, this proportion also can be refer as a adjusting parameter trade off between vertical and horizontal cooperation cost. The vertical cooperation cost increase through giving higher value of weighted parameter α (i.e. $\alpha \geq 1-\alpha$), on the other hand, it means reducing the horizontal cooperation cost, and promotes the across-chain horizontal cooperation of cluster supply chain, on the contrary, when $\alpha \leq 1-\alpha$ ($0.5 > \alpha > 0$), it amplifies the horizontal cooperation cost, which leads to disrupting coordination between one single chain and another one.

Furthermore, for when α belong to the area $\alpha \geq 1-\alpha$ ($0.5 \leq \alpha \leq 1$), it also implies that the vertical cooperation channel is a regular and long term strategic channel, while across-chain horizontal cooperation is temporary and supplementary channel, for the cost incurred in strategic channel is lower than that in temporary channel, it matches with real situation occurred in business world. In this way, the paper will put more emphasis on α belonging to the area $0.5 \leq \alpha \leq 1$to explore the change of cluster supply chain design.

VI. CONCLUSION

We have presented a novel framework for intelligent model design cluster supply chain in which there are a multiple of rivals or potential competitors in the proximity for each member along supply chain. It means industrial cluster not only contains a couple of focal firms locating at the same tier, but includes the corresponding upstream and downstream firms as well, all of which concentrate on a close geographical site. Thus, it is most likely to form multiple paralleled single supply chains for each focal firm of industrial cluster, these paralleled single supply chains compete and cooperate with each other, that is to say that these single supply chains led by each individual focal firm have interrelated or intertwined each other less or more.

REFERENCES

[1] Christopher, M. (2005). *Logistics & Supply Chain Management*. FT Prentice-Hall, Harlow.

[2] Fletcher, M., Brennan, R.W., & Norrie, D. H. (2002). Modeling and reconfiguring intelligent holonic manufacturing systems with Internet-based mobile agents. *Journal of Intelligent Manufacturing*, 14: 7-23

[3] Kaufman, L., & Rousseeuw, P.J. (1990). *Finding Groups in Data: Introduction to Cluster Analysis.* Wiley, New York, NY.

[4] Pandit, N. R., Cook, G. A., & Swann G. M. P. (2002). A comparison of clustering dynamics in the British broadcasting and financial services industries. *International Journal of the Economics of Business*, 9(2):195-224.

[5] Punj, G. & Stewart, D. (1983). Cluster analysis in marketing research: review and suggestions for application. *Journal of Marketing Research*.20: 134-148.

[6] Shen Z. J M., & Qi L. (2007). Incorporating inventory and routing costs in strategic location models. *European Journal of Operation Research*, 179:372-389.

[7] Shen, Z. J. M. (2005). A multi-commodity supply chain design problem. *IIE Transactions*, 37: 753-762.

[8] Tragantalerngsak, S., Holt, J., & Ronnqvist, M. (2000). An exact method for the two-echelon, single-source, capacitated facility location problem. *European Journal of Operation Research*, 123(8): 473-489

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 427-430

The Application of Information Visualization in Business Site of China

Feng Yang
Guangdong Province Key LAB of Electronic Commerce Market Application Technology, Guangdong University of Business Study, Guangzhou, China
Email: fyang68@163.com

***Abstract*—Various techniques have been researched after the information visualization conception was proposed at 90th 20 century. These techniques come from different fields and have many styles. There are many kinds of information in E-Commerce systems. The application Information visualization in E-Commerce is certain. This paper analyzed the kinds of information in E-Commerce systems, chose the information that is appropriate for visualization, gave some examples and discussed the statues of information visualization application in E-Commerce.**

***Index Terms*—Information visualization, E-Commerc, Information management**

I. INTRODUCTION

Electronic Commerce is a kind of business running on Internet. Almost all new ideas, techniques and methods may be applied in it. From the view of information management and information system, they all can improve electronic commerce system.

As a new technique of information management, information Visualization will definitely be used into electron business system. In this paper, we research the application of information visualization in E-commerce. Information Visualization is one kind of technology expressing complicated relations between large amount of information using visual method such as images, animated cartoons [1]. It is becoming the new research hot that we find the implicit law from large amount of finance and commerce data and providing the basis for decision. For the consumer needs information visualization expresses the external and inner relation among the information collections by various appropriate visualization signs. We can discover knowledge concealing in information conveniently and promptly [2].

The different kinds of information, the complicated application of information field and various needs of consumers lead to people study and develop different forms of visualization technology. There are many kinds of information in electron business system. The different visualization forms can be used for the same kind of information and a certain visualization form can be used for many information sets.

In this paper, we firstly analyzed the information kind in electronic commerce and found which kind of information management can be visualized. Then we concluded the simple forms of visualization in e-commerce system. At last, we analyzed the situation of applications.

II. THE VISUALIZATION OBJECTS IN E-COMMERCE

From information management and the information system view, the operation of a successful electronic commerce system must be supported by a perfect information system platform. Information was stored in the system by data form. The information management was expressed as data management. Information management includes some aspect such as information collecting, organizing, searching, communicating and making use of information. Information organization is basis of information resources administration. A fine information processing means is the key of information management.

There are two kinds of information related to e-commerce system. The can be divided into five situations.

(1) One kind is entity information, includes commodity news, customer information, merchant information.

(2) The other kind is information connecting between them. They are buying recorders (these recorders reflected shopping behavior of customers, the extent of selling goods etc.), supplying recorders (reflecting buying ability of merchant)

Generally speaking, there are no direct relationship between the customers and the merchants.

Some technologies are targeted at one particular type data in information management and procession. Such as network navigation techniques based on organize commodity. Recommended information technology gets customer classification mainly based on customer information and buying recorders.

As a new information management technology it is inevitable that information visualization technology is the gradually applied in e-commerce. The use of information visualization techniques can mainly be around the above-mentioned types of information.

The first thing is definition the objects in Information visualization, Then we must determain which aspects of visual objects will be visualized. Finally, we should select the appropriate visual symbols and algorithms to complete the process. In this article we discussed the first two questions of information visualization in e-commerce.

Information visualization is an information technology that represents the wealth of information and the complex

relationship between them use visual symbols. In e-commerce system there are 3 kinds of information and 2 kinds of relationship will be visualized. At present the information in the e-commerce system visualization is mainly based on "commodity", "merchant" and "supply" information. It is now unusual for other two kinds of information.

The purpose of information visualization is finding the characteristics and laws hidden in the information by the interacting visual interface, causing people understanding information quickly and accurately. Describing the characteristics of things, laws, there are several different levels. We can use for a data. This is called as the concept visualization. We can also visual the relationship, such as classification and rules visualization. In the current e-commerce systems, we analysed their function and found that most of them were concept information visualization. Some of them were visualization of information classification and it was unusual for information visualization based on other forms.

III. TYPICAL APPLICATION IN E-COMMERCE INFORMATION VISUALIZATION

A. Concept-based information visualization

This is currently the largest form of information visualization. In this application visual objects mostly are "commodity" information and also involve some additional information.

1) Quasi-visualization of information characteristics

Generally, we describe the various features of things using numeric or non-numerical data, such as storing ex-factory date for every goods. Once the data sign distinguishing the old and new goods, of each commodity, they had an "old and new" features.

Almostly on every current interface of e-commerce systems we always give a commodity classification structure to help people find products, to achieve navigation. "Made in China net" is no exception. Behind the text of each sub-category there is a red color of the word "new" that shows the class of goods "old and new" level.

In This case, at least two places hadn't reached the standard of information visualization. Firstly it did not use graphical symbols fro feature expression. If we understood the red word "new" as a special icon, or with the appropriate graphical symbols instead, then the above examples seem to have information visualization functions. But the aim of information visualization is to reveal the essential characteristics of things, not to define concept for users. Different user groups have different standards about "old and new" in different periods. Information visualization should give a friendly graphical interface that a variety of users can intuitively understand the characteristics of things as far as possible. By this interface users can define their respective criteria related to their own tasks. Therefore, the example above has only some similar to information visualization to some extent.

2) Visualization of things characteristics

The characterization of merchant is visualized on "China E-Commerce Information Network". On its home page, we select the "business" and then click on "business, culture and education," a description of merchant information appears on the screen. Clicking on "printer", the information as figure 1 is shown. It visualizes the level of businessman description which is divided into four levels. The more red stars they have the higher level they are.

Figure 1. Businessman description in "China's Information Network E-Commerce"http://www.cneb.net(2009/06/26)

Figure 2. The description of goods characterization in "yahoo shopping" http://shopping.yahoo.com(2009/06/26)

Figure 3. the description of commodities class in "8848 site" http://www.8848.com(2009/06/26)

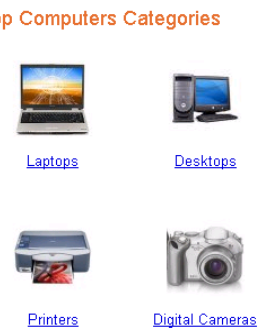


Figure 4. the description of commodities class in "yahoo shopping" http://shopping.yahoo.com(2009/06/26)

The same idea also appears in yahoo. The more sophisticated description using red star is adopted in

“yahoo's shopping”. For each product it is marked using five red and gray stars. The total number of stars shows the maximum possible level, the red star expresses the level of this product. In this way, not only the absolute characteristics of the goods is described but also expressed the relative characteristics.

3) The visualization of general characteristics

In the commodity classification structure, the last one is of course a specific commodity. Its features, such as quantity, the situation of new and old, can be visualized. For each node in the structure, it represents a class of goods. In addition to the above-mentioned similar problems, their general characteristics is also be visualized. The most typical method is for each classification we give o a classification symbol. Figure 5 shows the symbols of commodity classification in the “8848 site”. Figure 6 is the graphical signs used in yahoo shopping. Many commercial sites have adopted this method.

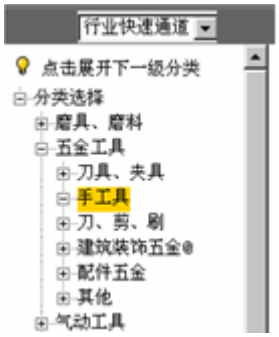


Figure 5. Commodity information tree in "HC net" http://mmt.machine.hc360.com/merchant/sale/003006003/1.htm(2009/06/26)

B. *Information visualization for hierarchical structure*

More meaningful information visualization forms are visualized the relationship between the various structures. If a system can help people understand the global structure and local features at mean time it may be accepted easily. Information Visualization is always trying to make it easily for users to grasp the overall situation and the details simultaneously. These forms of information relationship include tree, association rules, trends, clustering and deviation. There are many visualization technology based on these forms [3] [4] [5], however, few of them little is currently applied in e-commerce system.

The tree visualization technique was adopted in "HC Network" for product information organization. The system uses outline view. This technique comes from the earliest tree visualization technique contour map [6], it was used to visualize the disk's file structure. In UNIX system the "du" command also used this form. In Windows operating system it was designed more perfectly.

IV. Analysis of Application

The several typical e-commerce systems above basically covered all the current the application of information visualization technology in China. Information visualization application in e-commerce system is relatively backward compared with other information systems. In some information systems, information visualization technology already has more applications. For example, SPSS, DB-Miner and Oracle all adopted Information visualization techniques for expressing results. Some problems of information visualization exist in e-commerce systems of China.

A. *Narrow coverage of information*

Most information in these systems is "commodity", then the "Merchant". In fact information visualization can also be carried out for "customer", "shopping" and "supply", such as the customer level, shopping volume. Now we can’t find example. Some common characteristics which possibly can be visualized include:

(1)Customer information such as age, income status, geographical distribution;

(2)Commodity information such as number, inventory number, price etc.;

(3)Merchant information, includes geographic distribution, size, brand grade, social reputation, etc.;

(4)User interesting degree analyzed from shopping log, customer loyalty, customer-star, business integrity, sales speed, etc.;

(5)Merchants level analyzed from supplication availability, integrity level and so on.

B. *Monotonous visual form*

1) The tree structure

There are many information visualization techniques. These technologies can be divided into two categories: The first type is node connections visualization technology, includes H-tree, radial view, balloon view, hyperbolic-tree, cone-tree etc. [7] [8]. The second category is based on space-filling type of tree visualization techniques, such as tree-map, icicle-plot, tree-ring [9], and Cheops [10]. All of these technologies had been applied in other information system, however, in e-commerce system we can’t see any appearance.

2) The icons selection

When we visualized the tree structure of the information sets the important thing is expressing the relationship of the nodes in the hierarchy. One of the main approaches is to design a good map symbol. We should pay attention to the following:

(1) Intuitive icons selection.

The purpose of visualization is to enhance the people's cognitive abilities, improve understanding efficiency and reduce the complexity in using information. If the system selected the Non-intuitive icons they would increase the burden of understanding. In figure 5 it would be difficult to understand the meaning of icons without the text specifications.

(2) Language independence of icons

Because of different cultural backgrounds people speak different language. People generally can not be read in text written by other language but understand the same icons. In this way e-business system can be used by all countries easily. In figure 6, due to the rationality of icons, most users can understand the type of goods without text description.

(3)Standardization of icons

Building standardized icons base is the important step. The choice of the standard icons must conform to the habits of industries and local areas.

C. Low-level of visualization

Information is the description of external features and internal features of things. Description a thing with just one data is not enough. We always describe a thing with more than one data and call these data as properties of things. In order to describe a particular attribute of things we can use a simple data, such as values, character, an image, a sound and so on. For a complete description of things we need to use a composite data made up with some simple data, for example, records in the database technology.

Now it is still at a low level for information visualization applied in E-commerce. We can just find some features of information visualization. Information visualization of the relationship among the information sets only can be found a few examples.

V. FURTHER STUDY

From above analysis, we can find the application of information visualization techniques for E-commerce websites in China is still in the initial stage. Future applications can be from the following aspects.

First of all, we can increase product type icons. For some products there are the standard icons in traditional business, we can directly adopt them for their on-line product type signs. If there are no standard icons we need design new icons database for different industries and products.

For commodities navigation system we can design content-based hierarchical tree structure. Because the hierarchy information visualization have two classes as node-link style and space-filling style. They are also called connection style and enclosure style in some literature. E-commerce site is currently only used node-link-type. For some websites there are many products in them and the clients always need to compare the number of different categories, we can use space-filling type to visualize the relationship.

Network structure visualization can be used on the site map. The current site maps only list the function of the websites simply and don't show the relationship between them. The Information Visualization in this application will have more opportunities.

REFERENCES

[1] Zhou Ning, Wen Yanping & Liu Wei. On the methods of information resources visualization. The Proceedings of digital library-IT opportunities and challenges in the new millennium. July 2002

[2] Jiawei H, Kamber M. Data mining concepts and techniques. Morgan Kaufman Publishers, Inc. 2001. 230-300

[3] Jing Y, Ward MO, Rundensteiner EA. Interactive hierarchical displays: a general framework for visualization and exploration of large multivariate data sets. Computers & Graphics 27(2003). 265-283

[4] Chaomei C. Empirical evaluation of information visualizations: an introduction. Int. J. Human Computer Studies. Vol 53. 2000. 631-635

[5] Hujun Y. Nonlinear multidimensional data projection and visualization. Intelligent Data Engineering and Automated Learning 2003. Springer-Verlag Berlin Heidelberg 2003. ISBN 0302-9743. 377-388

[6] Johnson B, Shneiderman B. (1991) "Tree-Map: a space filling approach to the visualization of hierarchical information structures" IEEE Visualization '91. 1991. 284-291

[7] Herman,G. Melançon,M.S. Marshall. (2000) "Graph Visualization in Information Visualization: a Survey" IEEE Transactions on Visualization and Computer Graphics. 2000. 24-43.

[8] Nguyen,Q.V., Huang,M.L. (2002) "A Space-Optimized Tree Visualization" Proceedings of the IEEE Symposium on Information Visualization 2002 (InfoVis'02), 2002. 85-92

[9] Barlow,T., Neville,P. "A Comparison of 2-D Visualizations of Hierarchies" Proceedings of the IEEE Symposium on Information Visualization 2001 (InfoVis'01). 2001. 131-138

[10] Beaudoin,L., Parent,M.A., Vroomen,L.C. (1996) "Cheops: A compact explorer for complex hierarchies" Proc. Visualization'96. 87–92

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 431-434

Applying Association Rule Analysis in Bibliometric Analysis

——A Case Study in Data Mining

Fang Li, Chengyao Li, and Yangge Tian*(correspondence author)
International School of Software , Wuhan University, 129 Luoyu Road , Wuhan, China
tiandebox@126.com

***Abstract*—Scientific research needs lots of literature searches that cost a large amount of time and energy. Bibliometric techniques could help us find research hotspots and grasp the research direction yet can't reveal huge hidden information in massive literature since the existing bibliometric analysis techniques employ mainly simple statistical analysis techniques. Hence, we propose to introduce data mining analysis techniques into bibliometrics analysis, and expect to reach some instructive conclusions by mining relations among information like keywords, authors, research institutions, publications and so on. We take the subject 'data mining' as our research object and analyze the records by the method of association rule analysis. Finally, we get some valuable conclusions that verify the idea's feasibility. The results of this research would not only provide a reference for data mining research but also be applied to other research fields.**

***Index Terms*—Terms: data mining; bibliometric; association rule; SCI**

I. INTRODUCTION

Large amounts of literature always need to be queried during conducting scientific research, thus it is very crucial to quickly find out research hotspots and grasp the main development directions of current academic research from those documents. Manual analyses which mainly employs simply methods that are not only time-consuming and laborious but also heavily dependent on researcher's personal experience and research interests, often fail to fully extract implicit information and inherent laws.

Nowadays, more and more researchers begin to use biliometric techniques with statistical analysis of literature content information, citation information, author information, external features of documentation and other related information, so as to provide useful guidelines for research work[1][2]. However, existing bibliometric techniques, while emphasizing using mathematical and statistical methods to describe, evaluate and predict the status and development trend of science and technology, generally use relatively simple mathematical methods (which are basically elementary mathematics methods), derive relatively plain conclusions, and provide limited guidance for scientific research. Hence, we introduce data mining techniques into bibliometric methods.

Data Mining, also known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases [3]. With the rapid growth in a variety of data, data mining has become an important research topic and is receiving substantial interest from both academia and industry [4].

Association rule mining is one of the most active research directions in data mining. It reflects the interdependence and relevance between things. If there are certain associations between things, then one's information could be predicted by analyzing others'. An association rule problem between itemsets in mining customer transaction database was first proposed by Agrawal in 1993[5], and since then many researchers have conducted a lot of researches regarding mining problems of association rules in the future. At present the technology is already widely applied in business, medicine, earth sciences and other fields.

It is easy to discover some similarities between shopping basket data (which are commonly used in association analysis) and keywords, authors as well as other data sets in bibliometric analysis, so we consider using the application of association rules in bibliometric study and make use of such data mining methods to find hidden information and laws from a large number of the literature data.

II. METHODS AND MATERIAL

In bibliometric study, the Science Citation Index (SCI), provided by the Institute for Scientific Information (ISI) Web of Science databases, is the most important and frequently used source database for the review of scientific achievement in all research fields [6][7].

The data we used was extracted from the SCI online database. Data mining and its synonyms (such as data mine, KDD, exploratory data analysis, information discovery, information extraction, intelligent data analysis etc.) were used as the search topics. All the information was obtained on July, 16th, 2008 when the SCI search process for this study was conducted. The total number of papers related to data mining research in the ISI web database was 10286. After deleting the repeated records, the total number was 9808, which were published between 1962 and 2008. These were published with 13 document types with the distribution analysis. There were 8930 regular/research articles, which accounted for 91.07% of the total production, followed by reviews (303; 3.09%), meeting abstract (214; 2.18%), editorial materials (194; 1.97%), book review(78; 0.8%), letter (26; 0.27%),news

AP-PROC-CS-09CN005

items (14; 0.14%), note(14; 0.14%), correction (9; 0.09%), Biographical-Item (5; 0.05%)and reprint (2; 0.02%).

We extracted from SCI database key words, authors, source, time cited, author address and some other records as its original form, and then transformed them to formats that fit for association rule analysis. Finally, all these records were imported into SQL SERVER 2005.

III. METHODS

In our research, we primarily used data mining techniques and adopted the classical association rules algorithm in analysis. The definition and algorithm of association rule are introduced below:

A. Definition

Let $I = \{i_1, i_2, ..., i_m\}$ be a set of items. Let D, the task relevant data, be a set of database transactions where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$ and $A \cap B = \phi$. The rule $A \Rightarrow B$ holds in the transaction set D with support s, where s is the percentage of transaction in D that contain $A \cup B$. The rule $A \Rightarrow B$ has confidence c in the transaction set D if c is the percentage of transactions in D containing A which also contain B. That is,

$$\text{support}(\text{A} => \text{B}) = \text{Prob}(A \cup B)$$

$$\text{confidence}(\text{A} => \text{B}) = \text{Prob}(B \mid A)$$ [8].

Support for association rule reflects the frequency of the rule while confidence indicates the accuracy of the rules. In this study, we set the min_support as 5 and the min_confidence as 20%. (In some case, we set the min_support as 3 in order to reach better results.)

B. Algorithm

Association rule mining generally has two steps[8]:

Step 1: Find all frequent itemsets. By definition, each of these itemsets will occur at least as frequently as a pre-determined minimum support count.

Step 2: Generate strong association rules, which are bigger than or equal with the minimum support and minimum confidence, from the frequent itemsets.

During data mining, we use Apriori Algorithm which is the most influential algorithm for mining association rules[9]. Apriori employs an iterative approach, where k-itemsets are used to explore (k+1)-itemset. First, find the set of frequent 1-itemsets which is denoted L_1. L_1 is used to find L_2, the frequent 2-itemsets, which is used to find L_3, and so on, until no more frequent k-itemsets can be generated. The following two steps are responsible for the process of finding L_k through L_{k-1}:

The join step: C_k is generated by joining L_{k-1} with itself. C_k Here stands for Candidate itemsets of size k.

The prune step: C_k is a superset of L_k, that is, its members may or may not be frequent, but all of the frequent k-itemsets are included in C_k. Any (k-1)-itemsets that is not frequent cannot be a subset of a frequent k-itemsets, hence should be removed.

IV. RESULTS

Through this method, we get some interesting results. After analysis, we reached the following conclusions:

A. Keywords analysis

In traditional keywords analysis, we generally find research hotspots and research directions through statistics on the number and changes of papers' common keywords at different times, as shown in Table I.:

TABLE I. FREQUENCY OF AUTHOR KEYWORDS USED IN PUBLICATIONS—TOP 30(EXCERPT FROM REFERENCE[10])

DE	*1999-2008*	*P (%)*	*1999-2003*	*P (%)*	*2004-2008*	*P (%)*
data mining	2472	26.65	809	25.74	1663	27.11
clustering↑	279	3.01	68	2.16	211	3.44
classification↑	266	2.87	81	2.58	185	3.02
machine learning	254	2.74	97	3.09	157	2.56
Knowledge Discovery ↓	219	2.36	110	3.50	109	1.78
association rules ↑	206	2.22	73	2.32	133	2.17
information extraction↑	198	2.13	57	1.81	141	2.30
exploratory data analysis ↓	156	1.68	76	2.42	80	1.30
bioinformatics↑	146	1.57	36	1.15	110	1.79
neural networks ↓	136	1.47	62	1.97	74	1.21
decision trees↓	91	0.98	38	1.21	53	0.86
association rule↑	80	0.86	21	0.67	59	0.96
feature selection	78	0.84	24	0.76	54	0.88
decision tree	73	0.79	18	0.57	55	0.90

Yet, mere statistical analysis cannot fully reveal the implicit rules among keywords. Sometimes, some research directions were neglected by researchers simply because the authors chose different keywords. The association analysis can just make up for this deficiency by analyzing the correlations between keywords and thus presenting the relations between different research fields. As shown below:

TABLE II. ASSOCIATION RULES IN KEYWORDS

Confidence	*Rules*
1.000	motifs , patterns--> clusters
0.857	signal detection--> pharmacovigilance
0.833	Multidimensional scaling, feature extraction--> Sammon mapping
0.429	workflow management--> Petri nets
0.429	workflow management--> process mining
0.417	combinatorial chemistry--> nonlinear mapping
0.357	granular computing--> rough set
0.353	fuzzy set--> quantitative value
0.294	fuzzy set--> rough set
0.278	Web usage mining--> clustering
0.263	outlier detection--> clustering
0.250	discretization--> machine learning
0.250	discretization--> classification
0.240	decision trees--> classification
0.240	decision trees--> classification
0.238	entropy--> machine learning
0.231	unsupervised learning--> clustering
0.231	unsupervised learning--> clustering
0.227	regression--> classification
0.219	rule induction--> machine learning
0.219	rule induction--> machine learning
0.200	support vector machines--> machine learning
0.200	feature selection--> classification
0.200	self-organizing map--> clustering

Table III presents part of the important rules we discovered through association rules analysis, sorting them according to each rule's confidence. The majority of these rules indicate research methods or techniques in certain area, such as, 'multidimensional scaling, feature extraction -> Sammon mapping', 'workflow management -> Petri nets' and so on, while some rules show the correlations between two methods or techniques, like 'fuzzy set--> rough set'. These findings are of great instructive importance to researchers.

Compared to the conclusion reached by traditional bibliometric analysis, as shown in Table III, results of analysis employed association rules are apparently more informative and enlightening. For instance, from Table II, we can only figure out that clustering is an important research direction, yet we discovered that clustering is closely related with many techniques like outlier detection, unsupervised learning, self-organizing map and so on, and could be applied in motifs patterns and web usage mining analysis.

Though some association rules' confidence is relatively lower than others, it doesn't necessarily means that it is of little importance. Confidence is influenced by popularity. That is, if a keyword refers to a new technique which was new introduced, it's thus hard to be discussed widely. Consequently, rules associate with such keywords are not possible to be of high confidence yet still have great value in instructing research. For instance, the rule 'Web usage mining -> clustering' possess' has a low confidence since 'web usage mining' is a new application of data mining and appeared only recent years. This rule indicates that clustering methods could be applied in 'web usage mining'. Though it is not widely been discussed yet, it enlightens other researchers who are interested in this area.

B. Keywords and Journals

Similarly, we studied association rules between keywords and journals. As shown in Table IV, the analysis reflects different journals' preference of keyword. For example, ACM journals prefer data mining algorithms while International Journal on Document Analysis and Recognition shows particular interest in information extraction. These findings could lead researchers to read some journals correspond to their research interests, or choose the proper publications to submit their research results.

TABLE III. ASSOCIATION RULES BETWEEN KEYWORDS AND JOURNALS

Confidence	*Rules*
1.000	SURFACE AND INTERFACE ANALYSIS -> chemical analysis
1.000	ACM COMPUTING SURVEYS -> algorithms
0.944	ACM TRANSACTIONS ON DATABASE SYSTEMS -> algorithms
0.750	INTERNATIONAL JOURNAL ON DOCUMENT ANALYSIS AND RECOGNITION -> information extraction
0.750	APPLICATIONS OF BIOINFORMATICS IN CANCER DETECTION -> bioinformatics
0.737	INTERNATIONAL JOURNAL OF DATA MINING AND BIOINFORMATICS -> bioinformatics
0.667	JOURNAL OF COMPUTATIONAL CHEMISTRY -> combinatorial chemistry
0.600	JOURNAL OF STATISTICAL MECHANICS-THEORY AND EXPERIMENT -> data mining (experiment)
0.438	ACM TRANSACTIONS ON INFORMATION SYSTEMS -> algorithms
0.417	KYBERNETES -> cybernetics
0.375	ANNALS OF OPERATIONS RESEARCH -> classification
0.333	DATA WAREHOUSING AND KNOWLEDGE DISCOVERY, PROCEEDINGS -> association rules
0.300	IEEE TRANSACTIONS ON SOFTWARE ENGINEERING -> association rules
0.286	JOURNAL OF PROTEOME RESEARCH -> bioinformatics
0.273	ADVANCED DATA MINING AND APPLICATIONS, PROCEEDINGS -> association rules
0.263	IEEE TRANSACTIONS ON SYSTEMS MAN AND CYBERNETICS PART A-SYSTEMS AND HUMANS -> classification
0.231	AMERICAN STATISTICIAN -> EXPLORATORY DATA ANALYSIS

C. Authors and Keywords

The analysis of relations between authors and keywords revealed the foremost researchers in certain area. For example, as shown in Table VI, 'Petri nets, workflow management -> van der Aalst, WMP' suggests that the main researcher about Petri nets and workflow is van der

Aalst and WMP. Hence, other researchers interested in this area could pay more attention to his papers, publications and also research trend.

TABLE IV. ASSOCIATION RULES BETWEEN KEYWORDS AND AUTHORS

Confidence	*Rules*
1.000	clusters, patterns -> Parida, L
1.000	clusters, motifs -> Parida, L
1.000	nonlinear mapping, combinatorial chemistry -> Agrafiotis, DK
0.833	Petri nets, workflow management -> van der Aalst, WMP
0.833	quantitative value, fuzzy set -> Hong, TP
0.750	classification problems -> Hu, YC
0.714	information visualization, visual data mining -> Keim, DA
0.556	rough sets, pattern recognition -> Pal, SK
0.500	information fusion -> Wang, ZY
0.500	information fusion -> Leung, KS
0.444	soft computing -> Pal, SK
0.357	decision making -> Kusiak, A
0.316	data clustering -> Chen, MS
0.313	evolutionary computation -> Wong, ML
0.313	evolutionary computation -> Leung, KS
0.294	fuzzy set -> Wang, SL
0.263	Grid computing -> Talia, D
0.261	distributed data mining -> Kargupta, H
0.240	Bayesian networks -> Wong, ML
0.227	customer relationship manageme -> Van den Poel, D

D. Keywords and Research institutions

Similar to analysis of relations between authors and keywords, analysis of relations between keywords and research institutions, as shown in Table VII, indicates institutions' preference in some certain research area. This helps researchers find out relevant research institutions and carry out academic exchanges.

TABLE V.

Confidence	*Rules*
1.000	chemical analysis -> Tottori Univ
1.000	biomedical literature data min -> Norwegian Univ Sci & Technol
1.000	Petri nets -> Eindhoven Univ Technol
1.000	CLASSIFIER DESIGN -> Ben Gurion Univ Negev
1.000	cascade generalization -> Univ Wisconsin
1.000	Choquet integrals -> Chinese Univ Hong Kong
0.750	information visualization -> Univ Konstanz
0.750	decision making -> Univ Iowa
0.600	cybernetics -> Portland State Univ
0.500	clusters -> Univ Haifa
0.500	classification problems -> Chung Yuan Christian Univ
0.500	distributed data mining -> Univ Calabria
0.333	data cleaning -> Florida Atlantic Univ
0.250	combinatorial chemistry -> 3 Dimens Pharmaceut Inc

Besides, we also conducted research on the relations between authors and journals, keywords and time cited and so forth, and came to some interesting results as well.

V. CONCLUSION

Association analysis is an important data mining technique, widely used in commercial, financial, telecommunications, medical fields and so on, but rarely applied in the bibliometric analysis. Our research shows that association rules can discover information hidden in the keywords, publications, authors, research institutions and other materials. In particular, it can instruct researchers to find research fields and techniques related to its research direction. At the same time, it helps them broaden research ideas and even discover new research fields by providing relevant publications, authors and research institutions. This has significant instructive value to research works.

Different from general rules, when applying association rules in bibliometrics, we should also pay attention to those rules with relative low confidence because scientific research stressed on innovativeness. Generally, new research directions have a relatively low confidence since researches that focus on them are limited yet. On the other hand, some rules with a relatively high confidence might have little guiding significance. Hence, making the right judgment on this matter requires researchers' research experience.

On the whole, the association rule analysis provides us a new, objective and credible approach to analyze and evaluate scientific literature. It instructs us conducting scientific research by quickly discovering implicit information and rules among numerous literature data and shows considerable research perspectives and application value.

ACKNOWLEDGMENT

This paper is supported by NSFC (Granted No. 40601026).

REFERENCES

[1] Cronin, 2001 B. Cronin, "Bibliometrics and beyond: Some thoughts on web-based citation analysis," Journal of Information Science, vol. 27, pp. 1–7, January 2001.

[2] H.F. Moed, R.E. Debruin and T.N. Vanleeuwen, "New bibliometric tools for the assessment of national research performance—Database description, overview of indicators and first applications," Scientometrics, vol. 33, pp. 381–422, July 1995.

[3] W. Frawley and G. Piatetsky-Shapiro and C. Matheus. "Knowledge Discovery in Databases: An Overview". AI Magazine: pp. 213–228, Fall 1992.

[4] JF Roddick, M Spiliopoulou. "A bibliography of temporal, spatial and spatio-temporal data mining research". ACM SIGKDD Explorations Newsletter, vol. 1(1), 1999, pp. 34 – 38.

[5] W.T. Chiu and Y.S. Ho, "Bibliometric analysis of homeopathy research during the period of 1991 to 2003," Scientometrics, vol. 63, pp. 3–23, March 2005.

[6] A.E. Bayer and J. Folger, "Some correlates of a citation measure of productivity in science", Sociology of Education, vol. 39 (4) , 1966, pp. 381–390.

[7] R.N. Kostoff, "The underpublishing of science and technology results", The Scientist, vol. 14 (9), 2000, p. 6.

[8] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, 2000, chap. 6 pp.4-5

[9] http://en.wikipedia.org/wiki/Apriori_algorithm

[10] Shuang Deng, Yangge Tian, "Using the bibliometric analysis to evaluate global scientific production of data mining papers"

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 435-438

Research and Realization of Complex Three Dimensional Stratum Modeling

Ning Zhao [1], Qian Zhan [2], and Weifeng Du [3*]
[1] School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China
Email: zhaoning@hpu.edu.cn
[2] The People's Bank of China, Jiaozuo Central Sub-branch, Jiaozuo, China
Email: qianqian82950@163.com
[3] The School of Mathematics & Information Engineering, Jiaxing University, Jiaxing, China, 314001
Email:woodmud@tom.com

***Abstract*— With the development of computer technology and geological modeling study of the gradual maturity, 3D for the engineering geological modeling and information visualization analysis system of opportunities and challenges brought about. Have carried on the detailed introduction to the three dimensional geological modeling tool and the method. Summarized the three dimensional geological modeling basic step. Surface topography information deduce the initial construction of the geological model project . Carry out the calculation of occurrence. How to use the discrete value of the occurrence of information, production level surface equation method , And study of the surface intersection, Initial generation of 3D geological model .This paper discusses the adoption of the new access to the geological information on how to modify the initial model editor .Application algorithm for the generation of surface fitting interpolation. Display system for the section of geological strata, faultage and three-dimensional geological model of the effect drawings of block geological. At the same time, it provides a reliable scientific authority for the decision-making of the underground construction projects.**

***Index Terms*—2D modeling, 3D modeling , block track Delaunay Triangulation**

I. INTRODUCTION

Three-dimensional engineering geological modeling refers to the proper establishment of the geological features of the data structure mathematical model and use computer graphics technology to mathematical description of the geological 3D photorealistic images to be expressed in the form. Not only can the use of 3D modeling technology and intuitive description of the complex underground geological conditions, the image to express the morphological characteristics of geological structure, as well as structural elements of spatial relations, and the combination of engineering geological information, can make analysis more intuitive and accurate, so as to fast, dynamic three-dimensional reproduction of engineering geology and geological information to develop a comprehensive analysis and effective way.

* Corresponding author：Du Weifeng , School of Mathematics & Information Engineering, Jiaxing University, Jiaxing, Zhejiang, China, Email:woodmud@tom.com

This modeling system is the source of geological information database based on multi-service support platform for the effective use of topographical, geological boundaries, faults, geological occurrence and small-surface histogram and other information on the establishment of surface geological model, the application borehole data, surface occurrence data extrapolating the initial geological vertical section. Since then, the application system provided by the longitudinal profile editor, editing by geological experts to amend the initial longitudinal section in the longitudinal profile with the correct, based on the generated cross-sections along the baseline.

II. THREE-DIMENSIONAL GEOLOGICAL MODELING

A. The definition of 3D geological modeling

We have the concept of two-dimensional block structure model can be extended to three-dimensional description of three-dimensional block structure model. Is not difficult to believe that three-dimensional block structure model that can actually be considered the same stratigraphic section in the adjacent block connections. Shall also belong to the same strata in the adjacent section of the envelope body. Each layer of blocks is a separate geological blocks, each a separate block with its own geological geological properties and the outer surface. The use of surface space for three-dimensional geological model describes the block can be three-dimensional geological block is defined as: $g = H_u \bigcup H_d \bigcup S_b$. Of which: H_u Block for the upper layer interface; Block for the lower layer interface; S_b To block surrounded by the lower layer interface for the closed boundary surface. Three-dimensional geological block gives the definition of the airspace defined three-dimensional geological model and the geological nature of the bulk of the set operations:

(1) Complicated geological structure model body formation can be three-dimensional reconstruction of geological blocks of space *n* independent body, and set descriptions, denoted by $G = \bigcup_{i=1}^{n} g_i$. This means: any complex three-dimensional geological model can be

applied to simple structure, a collection of geological unit body composition.

(2) In any complicated geological structure, the various geological block independent existence and disjoint, ie $\bigcap_{i=1}^{n} g_i = \phi$.

(3) Complicated geological stratigraphic structure model can be arbitrarily broken down into different numbers of sub-set of the geological block model, namely: $V = G - \bigcup_{i=1}^{l} g_i$

B. The Implement structure of 3D geological modeling

Three-dimensional geological block constituted by a number of facets, each facet has its own border curve and control points, where the boundary curve is a patch with other facets of the intersection line determines the surface film scope, while the control point determines the appearance of patches of the geometry. Three-dimensional surface of these films do not have the appearance of the rules are often complex, often use triangulation to represent the space of these complex surface film, will be a continuous three-dimensional discrete surface patches into a series of triangular mesh can be effectively expressed the three-dimensional geometric surface patches appearance. It is shown in Figure 1:

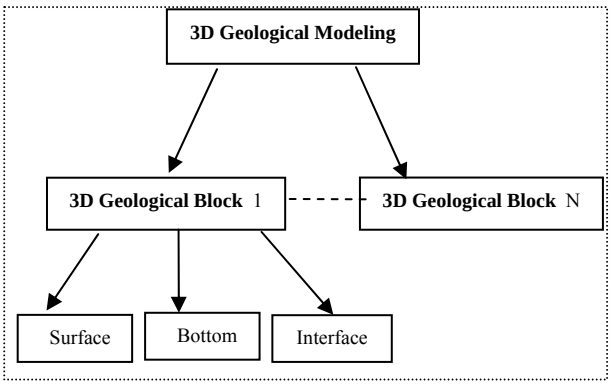


Figure 1. Three-dimensional geological block diagram topology

III. BASED ON THE THREE-DIMENSIONAL GEOLOGICAL MODELING PROFILE

A. The principle of the profile geological modeling

Three-dimensional geological block model is defined as a complex three-dimensional geological modeling to provide the basis of modeling method. As a result, the complex three-dimensional geological modeling into two adjacent geological strata of the two-dimensional cross-section block with the same body connection, for one stratigraphic section and the other a cross-section there is no such attribute for formation of the stratigraphic pinch-out processing block join algorithm research, namely through the strata block join algorithm processing block after the election of the three-dimensional structure of the geological unit body is enclosed envelope volume is a polyhedron composed of four. Figure in the C1, C2, C3 and C4, respectively, two-dimensional block stratigraphic structure, in which C2 for the pinch-out formation.

Is not difficult to believe that, due to changes in geological conditions, the complexity and diversity, the same connection will be very complicated stratum. But can be summarized as follows to connect several stratigraphic correlation:

(1) 1 to 1 relationship
(2) 0 to 1 or 1 to 0
(3) 1 or more pairs of a multi-relationship
(4) many to many relationship

B. The traditional method of 3D modeling

A simple three-dimensional geological building block approach is the use of two-dimensional geological block contour reconstruction three-dimensional shape. Achieve the two-dimensional geological block adjacent contour lines between the three-dimensional model reconstruction is to use a range of interconnected triangular piece will be two adjacent contour lines to connect in space. show in Figure 2

However, how to ensure connecting the three-dimensional surface model is reasonable, and has a good nature need to be carefully studied. Connecting adjacent two-dimensional geological block the various control points of the contour line formed by a number of basic triangular face, should constitute the interconnected three-dimensional surface, but not between the internal intersection of triangulated surfaces.

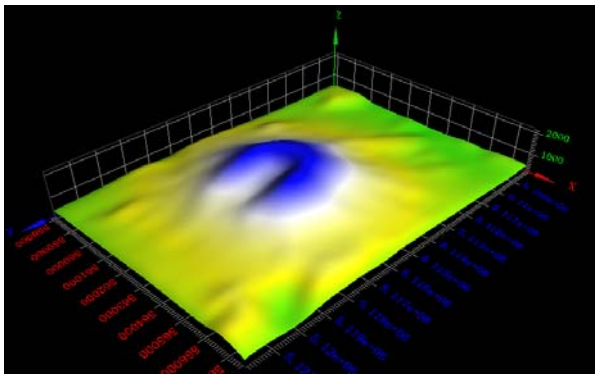


Figure 2. The traditional method of 3D modeling

IV. TRIANGULATION AND LIMITED ENCRYPTION ALGORITHM

A. Llimit the triangulation

Triangulation is to describe the basic elements of three-dimensional body surface is an important tool for three-dimensional geological modeling. The use of triangulation techniques can be drawn in a number of user profiles on the irregular distribution of stratigraphic control point to connect the triangular structure into a continuous surface to approximate the three-dimensional geological body surface. For a given n points $p_1, p_2, ..., p_n$, triangulation refers to disjoint line segment *connecting* p_i *and* p_j, $1 \le i, j \le n$, $i \ne j$, and so that each net is a

triangular area. Triangle Network is a plan, it has *n* vertices, it contains at most 3 *n* -6 edges. If you can give them the edge of a table, then get a solution of the problem. However, in practical applications often required for the triangulation made a number of constraints, this chapter in the Delaunay triangulation on the basis of further information on how to optimize three-dimensional subdivision.

B. Triangulation encryption algorithm

In a three-dimensional work area, the use of limited Triangulation Triangulation model can be established in the performance of the basic geometry of geological surface features, but if the user is given control points more sparse, or non-uniform density of data points, then constructed a triangular network a rough, partial rapid change. In order to construct a more smooth and delicate three-dimensional geological model, we must take a complete closed triangular mesh into several surface patches, in order to meet the requirements of precision surface patches, the need for further breakdown of surface patches. it is shown in Figure 3

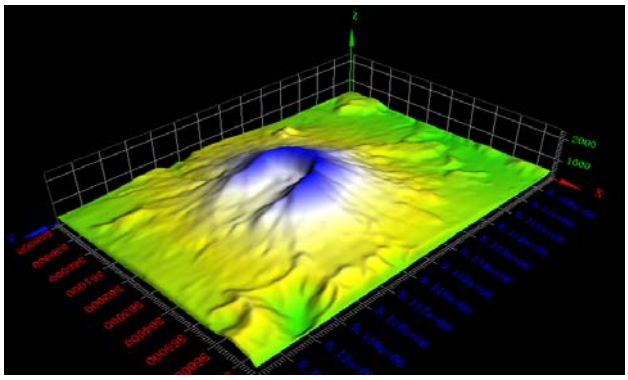


Figure 3. The Triangulation encryption algorithm of 3D modeling

According to computer graphics coordinates related to the concept of translation and rotation transformation, computing the first r blocks (j = 1,2,3, ... L) data on the total transformation matrix:

$$T_{m_j} = T_c \cdot R_x \cdot R_y \quad (4\text{-}1)$$

With the general transformation matrix, we can discrete points.

$$\left[x_r^{'}, y_r^{'}, z_r^{'}, 1\right] = T_{m_j} . \left[x_r, y_r, z_r, 1\right]^T \quad (4\text{-}2)$$

Spatial discrete data points after this transformation, and projected onto the corresponding plane, you can directly through the two-dimensional triangulation procedures to deal with. This block generates the triangular mesh over a simple combination is that we need a basic triangular grid. This basic right triangle mesh is subject to elimination of duplicate vertices, grid optimization, trajectory line generation process, three-dimensional discrete data can be three-dimensional subdivision.

C. Surface adaptive optimization rule

Adaptive subdivision surfaces is divided into sub-standard and sub-rules of two parts: a number of sub-criteria are used to control the subdivision process; subdivision rules used to recursively subdivided triangle. Subdivision standards are typically used to represent the maximum curvature surface. Directly solve the maximum surface curvature calculation of a large amount of practical application is a multi-curvature estimate.

Curvature Estimation of a variety of methods, such as the surface point to its planar approximation. shown in Figure 4. The maximum distance piece, or the height of the surface area divided by the surface to represent the other. the triangle vertices in the corresponding points in the outer surface method vectors N1, N2, N3, each vector of foreign law must be satisfied (1-Ni.Nj) <ε, i ≠ j and i, j ∈ (1,2,3). ε is tolerance, used to adjust the grid density. For the high-curvature regions are always segments continue, such as the Regional shrink to one point, then stop segments. In order to make triangular pieces as much as possible close to the surface to generate a better grid, ε generally taken to be 1.0 × 10-2-1.0 × 10-4.

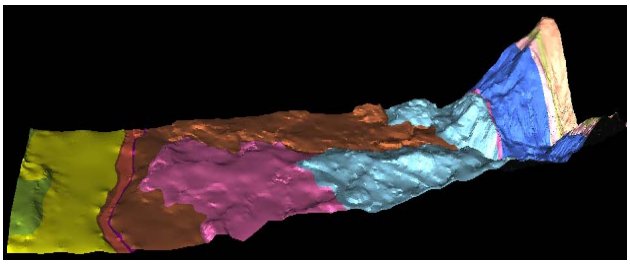


Figure 4. Surface adaptive optimization rule

Generate the initial triangular mesh surface, the surface in order to meet the precision requirements, the need for their networks to generate the curvature of each triangle test. Right triangle does not meet the tolerance requirements of a breakdown, while the relationship between changes related to the adjacent triangle until all triangles satisfy the tolerance requirements

IV. SYSTEM TESTING

Three-dimensional geological modeling function modules to achieve the main idea is: in the geological profile model based on the application described in

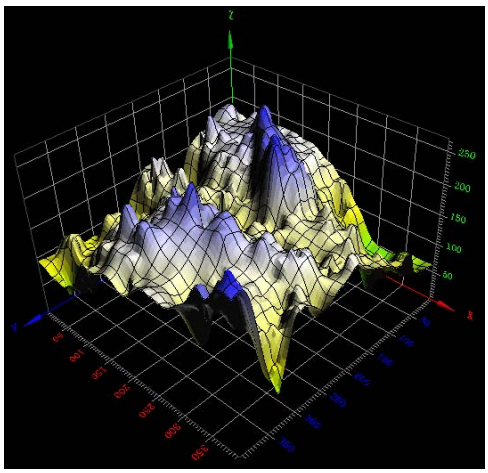


Figure 5. Testing A of 3D modeling

chapter IV of geological modeling principles, the interconnection between adjacent cross-section of the strata building block method of modeling three-dimensional geological model . Three-dimensional geological model of the formation contains a three-dimensional structure, and fault structure of the geological structure information. It is shown in Figure 5

For initial mesh scale, it can rectangular domain horizontal longitudinal ratio and will short side grid number set 1 or a base n, long side grid number Make length ratio and short side grid Number product . it is shown in Figure 6

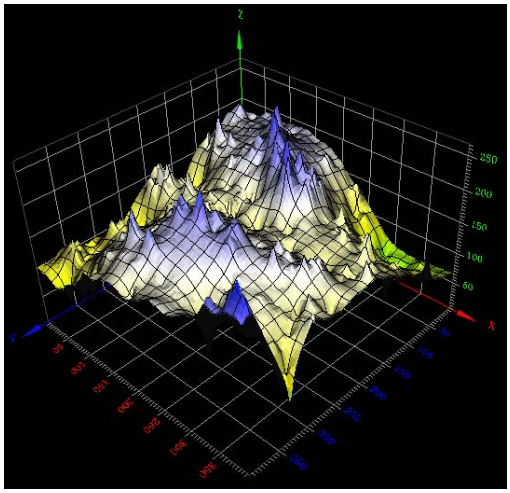


Figure 6. Testing B of 3D modeling

As is shown above, the whole performance of the system improve in more obvious results, reducing the appearance of the triangle tip-off.

B Test Conclusion

Through a series of comparison tests, adaptive Interpolation method for large-scale data is an effective solution to containing a large number of discrete data points of the surface surface interpolation.

V. CONCLUSION

We can draw the following conclusions through testing:

Using Based on the three-dimensional geological modeling profile has the following advantages: Application algorithm for the generation of surface fitting interpolation. Display system for the section of geological strata 、 faultage and three-dimensional geological model of the effect drawings of block geological. At the same time, it provides a reliable scientific authority for the decision-making of the underground construction projects etc.

ACKNOWLEDGMENTS

This work was partially supported by Zhejiang province fatal project (priority subjects) key industrial project (2008C11011).

REFERENCES

[1] Hestholm, S. O., Ruud B. O. 2-D Finite-difference elastic wave modeling including surface topography [J]. Geophys Prosp, 2004, 42: 371-390.

[2] Cerveny, V. Seismic rays and ray intensities in inhomogeneous anisotropic media[J]. Geophys. J. R. Astron. Soc, 2006, 29: 1-13.

[3] Guo B. Surface reconstruction from points to spline [J]. Computer Aided Design, 2007, 29(4): 269-277.

[4] Depreciation: Panel evidence for the G7 and 8 Latin American Countries. April.2005.

[5] Go Yonezawa, Tatsuya Nemoto, Shinji Masumoto et al. 3D Geologic Modeling and Visualization of Faulted Structures: Theory and GIS Application [C]. In: Proceeding of the Open GIS-GRASS users conference 2002-Trento, Itraly, 2002-09.

[6] Moser, T.J. Shortest path calculation of seismic rays [J]. Geophysics.2001, 56(1): 58-67.

[7] Angel, E. Interactive Computer Graphics: A Top-Down Approach with OpenGL, 3e. Edward Angel, 2002.

[8] American National Standards Institute, American National Standard for Information Processing Systems-Computer Graphics-Graphical Kernel System Functional Description, ANSI, X3.124-1985, ANSI, New York, 1985.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 439-442

Dynamic Research on a Water Walking Robot Inspired by Water Striders

Lan Wang [1], Tiehong Gao [1*], Feng Gao [2], Lina Dong [1],and Junnan Wu [1]
[1]School of Mechanical Engineering, Hebei University of Technology, Tianjin, China
Email: gaotiehong111@163.com
[2] State Key Laboratory of Mechanical System and Vibration, Shanghai Jiaotong University, Shanghai, China
Email: gaofeng@sjtu.edu.cn

***Abstract*—The superior speed and agility of water strider inspired researchers to design bionic robot walking on water surface. This paper proposed a novel buoyancy based water strider robot driven by decoupled parallel mechanism. Dynamic model of driving mechanism and hydrodynamic model of whole robot were developed in detail. Moving forward and turning experiments were done on prototype under the circumstance of laboratory. For its high efficiency, good mobility, low noise and little disturbance to environment, when equipped with a camera, this robot will have a great advantage in environmental monitoring and military reconnaissance.**

***Index Terms*—water strider, bionic robot, driving mechanism, dynamics, hydrodynamics**

I. INTRODUCTION

The living beings in nature are almost perfect through billion years of evolution. They give us inspiration to design biomimic robots working in unstructured circumstances.

Water strider can stand effortlessly and slide or jump quickly on water surface and rarely get wetted. Even on water as shallow as a tenth of an inch it still can walk. People become interested in water strider inspired robot walking on surface. Water strider robot can move fast efficiently and quietly using a sculling motion. So it is simpler and quieter than propeller. Such impressive advantages ensure water strider robot have a wide variety of applications from military equipments to civil devices, e.g. military reconnaissance, environmental monitoring, water quality examination, underground pipe detection, etc.

Various mechanical water striders have been developed in recent years. John Bush from MIT built the first water strider robot which was powered by an elastic thread and very similar to real insect in structure and motion mechanism[1], Metin Sitti from CMU developed STRIDE which used motors and Li-ion battery[2],Japan's Chuo University designed a buoyancy based water strider robot actuated by motors[3],and so on.

Excellent bionic robot design needs not only imitation but also innovation. Very light water strider keeps afloat by surface tension is ok, while for water stride robot it's not necessarily the same. To fulfill certain job on far-water surface independently, robot needs carry necessary electronics, actuators, power, sensors, and such. So the weight will be a great limitation to practical application[4,5]. With the development of material science, maybe in the future we can build a perfect water strider robot in various aspects, but not now[6,7]. So we developed a buoyancy based robot actuated by micro electromagnets using simple binary logistic control instead of complex servo control of motor. The robot moves and turns by sculling through decoupled parallel driving mechanism. Its simplicity, utility, agility and high load capacity can satisfy the demands of water strider robot.

In this article we will develop the dynamic model of driving mechanism of the robot. Considering water strider robot working on water, we'll also develop the hydrodynamic model of the driving mechanism and the whole robot. Finally we shall do moving, turning and loading experiments on prototype under the circumstance of laboratory.

II. ROBOT DESIGN

Water strider moves forward by rowing its two middle legs through an elliptic motion trajectory. To get the elliptic motion trajectory a 3 DOF decoupled parallel mechanism is designed as shown in Fig. 1.

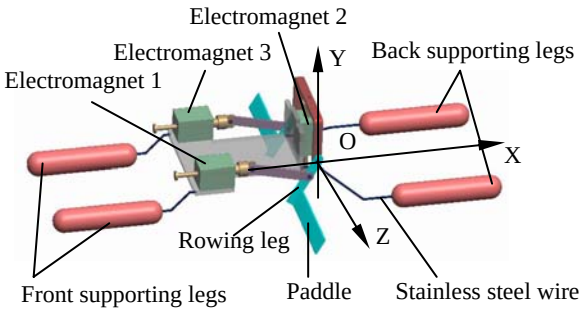


Figure 1 The virtual prototype model of the robot

The robot is actuated by three micro-electromagnets. Electromagnet 1 and 3 are used to achieve forward or backward motion of the driving mechanism, and the electromagnet 2 is for uplifting motion. The end point of rowing leg can get desired elliptic motion trajectory through cooperative control of the two electromagnets. Compare with motor actuated water strider robot, the

*corresponding author: Gao Tie-hong
National Natural Science Foundation of China (Grant No:30770538)

robot actuated by electromagnet have been simplified both in structure and in control method[8,9,10].

III. ESTABLISHMENT OF DYNAMIC MODEL OF ROBOT

Dynamic model of driving mechanism is built according to kinetic energy theorem. Based on fluid mechanics theory, hydrodynamic equations are formed for driving mechanism and robot.

Take account of the size of single leg driving mechanism is small, the weight of chosen material is much smaller than electromagnet output force, we suppose the friction of slider and revolute joint is zero in calculation. And so do the moment of inertia and the inertia force.

A. Establishment of dynamic model for driving mechanism

Single leg driving mechanism is simplified as shown in Fig. 2. The sizes of all components in driving mechanism are known, mass and moment of inertia and the position of centroid are also known. Slider 1 and slider 2 are actuators. Driving force F_{q1} and F_{q2} are generated by electromagnet 1 and electromagnet 2 respectively. The specifications of the two electromagnets are the same and electromagnet 2 is shared by two legs, so $F_{q1} = 2F_{q2}$. M_Z is the working resistance torque. Dynamic equation will be built based upon the above assumptions. We want the paddling driving force and acceleration on point C and D under actuation force of F_{q1} and F_{q2}.

The increment of total kinetic energy of the mechanical system in a transient is equal to the sum of work of all extern force acted on the system at that moment according to the kinetic energy theorem. Take the potential energy as the work of extern force. Here the influence of potential is neglected because of the light material and the low weight of moving iron core of electromagnet 2.

The total kinetic energy is expressed as

$$E=m_1v_{S_1}^2/2+J_{S_1}\omega_1^2/2+m_2v_{S_2}^2/2+Js_2\omega_2^2/2+m_Bv_B^2/2+m_Av_A^2/2 \quad (1)$$

Where m_1, m_2, m_A, m_B is the mass of link 1, link 2, slider A and slider B respectively; v_{s1}, v_{s2}, v_A, v_B is the velocity of link 1, link 2, slider A and slider B respectively; ω_1, ω_2 is the angular velocity of link 1 and link 2 respectively; J_{s1} and J_{s2} is the moment of inertia of link 1 and link 2 respectively.

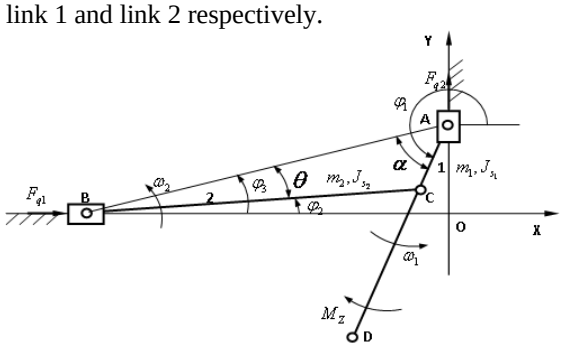


Figure 2 The dynamic analysis sketch map of the single leg mechanism

The increment of total kinetic energy is

$$dE=d(m_1v_{S_1}^2/2+J_{S_1}\omega_1^2/2+m_2v_{S_2}^2/2+Js_2\omega_2^2/2+m_Bv_B^2/2+m_Av_A^2/2) \quad (2)$$

According to the kinetic energy theorem, we have

$$dE = dW \quad (3)$$

Where dW is the work of all extern forces, it can be calculated in

$$dW = (F_{q1}v_B + F_{q2}v_A - M_Z\omega_1)dt \quad (4)$$

Then the dynamic equation can be described as:

$$\begin{aligned} &d(m_1v_{S_1}^2/2+J_{S_1}\omega_1^2/2+m_2v_{S_2}^2/2+Js_2\omega_2^2/2+m_Bv_B^2/2+m_Av_A^2/2) \\ &=(F_{q1}v_B+F_{q2}v_A-M_Z\omega_1)dt \end{aligned} \quad (5)$$

B. Hydrodynamic equation of driving mechanism

The main resistance of robot when walking on water comes from water resistance of rowing legs and supporting legs. As the water circumstance is very complex we suppose the robot works on calm water surface and only care the water resistance brought by static water in this paper. So we needn't take into account of waves, flowing of water, and wave forming resistance, etc.

1) Water resistance of paddle

When fluid flows bypass an object, the force on it can be decomposed into flow resistance and lift. The paddle of water strider robot is influenced by flow resistance and lift of water when paddling. The lift is neglected because it is very small. The flow resistance includes frictional resistance and shape resistance. The frictional resistance is shear stress formed by friction between fluid and body surface. The shape resistance, also known as pressure drag, is resistance caused by pressure difference between the back and the front of the body when fluid flow bypass. So objects are often streamlined to reduce the shape resistance in engineering.

The flow resistance on paddle can be calculated by

$$F_Z = C_P \frac{\rho U_0^2}{2} A_P \quad (6)$$

Where F_Z is the flow resistance on the paddle, C_P is flow resistance factor, U_0 is flow velocity which is equal to the rowing leg velocity relative to static water, ρ is density of fluid and for water it is 1.0×10^3 kg·m^{-3}, A_P is the projection area of paddle in the perpendicular direction of flow.

The flow resistance factor C_P is determined by Reynolds number, shape and surface roughness of the paddle. It is difficult to draw from the theoretical calculation and always determined by experiments. Since the paddle of robot is a thin flat panel and the aspect ratio is 6.25 the C_P is 1.2.

2) Water resistance of rowing leg

The flow resistance on paddle can be regarded as uniform load. As the paddle is perpendicular to the

rowing leg, the paddle can be regarded as a cantilever beam with fixed end at point D.

The water resistance on point D is

$$F_{DZ} = F_Z = C_P \frac{\rho U_0^2}{2} A_P \tag{7}$$

The uniform load of flow resistance on paddle is uniformly distributed along the length direction. It is calculated by

$$q = F_Z / L_P \tag{8}$$

Where L_P is the length of paddle.

The distributed load of flow resistance brings about resistance bending moment on point D of rowing leg. It is

$$M_{DZ} = \int_0^{L_P} qzdz = \frac{1}{2} qL_P{}^2 = \frac{1}{2} F_Z L_P = C_P \frac{\rho U_0{}^2}{4} A_P L_P \tag{9}$$

For the paddle is perpendicular to rowing leg, the moment of M_{DZ} to the rotary center of point A is zero. So the total water resistance moment to the rotary center of point A is

$$M_Z = F_{DZ} L_{AD} = C_P \frac{\rho U_0{}^2}{2} A_P L_{AD} \tag{10}$$

Where L_{AD} is the length of rowing leg.

3) Hydrodynamic equation of driving mechanism

Using (10) and (5), we get the hydrodynamic equation of single leg driving mechanism of robot. That is

$$\begin{aligned} &d\left\{\frac{\omega_1^2}{2}[m_1(\frac{v_{s1}}{\omega_1})^2 + J_{s1} + J_{s2}(\frac{\omega_2}{\omega_1})^2 + m_2(\frac{v_{s2}}{\omega_1})^2 + m_B(\frac{v_B}{\omega_1})^2 + m_B(\frac{v_A}{\omega_1})^2]\right\} \\ &= \omega_1[F_{q1}\frac{v_B}{\omega_1} + F_{q2}\frac{v_A}{\omega_1} - C_P\frac{\rho U_0^2}{2}A_P L_{AD}]dt \end{aligned} \tag{11}$$

Equation (11) is equivalent forms of dynamic model. It describes the dynamic relation of single leg driving mechanism under the driving force of electromagnet and water resistance.

C. Hydrodynamic equation of robot

1) Water resistance of supporting leg

During the moving process the overall water resistance on supporting leg is composed of frictional resistance caused by under-water part of supporting leg and shape resistance. This resistance of supporting leg makes resistance to robot when walking on water. Refer to Eq. (6),we get the overall resistance on supporting leg

$$R_S = C_S \frac{\rho U^2}{2} A_S \tag{12}$$

Where R_S is the flow resistance on the supporting leg, C_S is flow resistance factor, U is flow velocity ,which is robot velocity on water surface here, ρ is density of fluid, for water it is 1.0×10^3 kg·m^{-3}, A_S is the projection area of supporting leg in the perpendicular direction of flow.

The flow resistance factor C_S is related to the shape of supporting leg. Different shape has different C_S . Since the supporting leg of robot is of cylindrical shape and its ratio of Length and Diameter equals to 5 so the C_S is 0.85.

2) Hydrodynamic equation of robot

During paddling the robot is temporarily considered as a system of particles. The total mass is M . In the vertical direction the gravity and the buoyancy are equal and opposite in direction, So the robot receives a zero force in this direction. We only study the forces in the direction of moving forward.

When robot rows across the water, the water resistance on paddle F_Z propels the robot moving forward, while the water resistance on supporting leg R_S prevents the robot from moving forward. Other resistance such as wave resistance, resistance produced on paddle when the paddle just thrusting into water or out of water is not taken into account in this paper.

Since the robot has two paddles and four supporting legs, according to Newton second law, we get the dynamic equation of robot

$$Ma = 2F_Z - 4R_S \tag{13}$$

Where a is the robot acceleration relative to ground.

Using (6) , (12) and (13) we get

$$Ma = C_P \rho U_0{}^2 A_P - 2C_S \rho U^2 A_S \tag{14}$$

Let $a = \dot{U}$, $U_0 = U - v_{Dx}$

Where v_{Dx} is the speed of point D on rowing leg relative to robot.

Then we get the hydrodynamic equation of robot moving on water surface

$$M\dot{U} = C_P \rho (U - v_{Dx})^2 A_P - 2C_S \rho U^2 A_S \tag{15}$$

Ⅴ. EXPERIMENTS ON PROTOTYPE

Based on the dynamic research and other previous job a water strider robot prototype was made. And we did moving, turning and loading experiments on the prototype in homemade tank.

A. Moving forward experiments

Moving speed changes with the size of supporting leg and paddle, driving frequency of electromagnet, load on robot, etc. When the supporting leg'radius is 23 mm and length is 100 mm, the paddle'length is 50 mm and width is 10 mm, driving frequency is 6.74 Hz, through putting different weight on robot, we get the changing moving speed, described as Fig. 3.

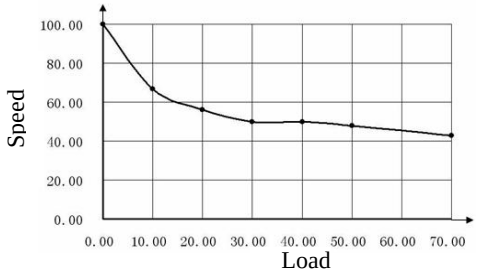


Figure 3 Moving forward experiments with load on robot

Under above conditions the highest speed of robot is 100 mm·s^{-1}.The robot slows down with the increasing load on it. If the load is under 40 g the speed is above 50 mm·s^{-1}.

B. Turning experiments

Turning experiments are done under the same condition as above except the paddle length is 60 mm. When the two rowing legs of robot are driven differentially, i.e., one rowing forwards and the other rowing backwards, the max angular speed is 12 deg·s^{-1}. The experiment results show the robot make a flexible turn with nearly zero radius when driven differentially. The error is mainly caused by the wire attached to the body of robot which prevents them from turning freely.

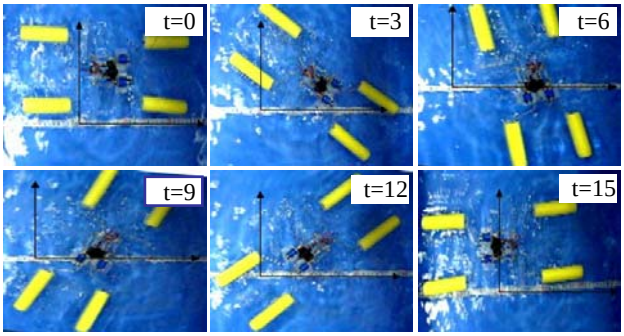


Figure 4 Photo snapshots of turning experiment driven at 6.74Hz

The forward speed is related to many factors from above experiments. Under certain range of frequency, robot moves faster with the higher frequency, longer rowing leg, shorter paddle, and lower load.

VI. Conclusions

According to the research on locomotion mechanism of water strider we put forward a six-leg water strider robot actuated by micro-electromagnet in this paper. Four supporting legs keep the robot from sink by buoyancy. The two rowing legs of robot have elliptic motion trajectory and drive the robot moving forward and turning just the same as its nature counterpart. We build the dynamic model of single leg driving mechanism. Further the hydrodynamic model is developed considering the robot works in water circumstance. Based on the dynamic research of whole robot a prototype is fabricated. Experiments are done on it under the circumstance of laboratory. The results show robot moving speed can reach 100 mm·s^{-1} and the peak turning angular speed is 12 deg·s^{-1}. Robot will perform better with optimized shape of legs and longer stroke of electromagnet. The developed water strider robot has the advantage of low cost, simple control system, high load capacity, practicality. Equipped with energy, improved control and detection system, the robot can work on far-water surface independently in future.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (Grant No:30770538),the Joint Research Fund for Overseas Natural Science of China(Grant No:50728503),and Doctoral Fund of Ministry of Education of China (Grant No:20060248030).

References

[1] David L. Hu, Brian Chan, John W.M. Bush. The design and construction of a mechanical water strider. *Elsevier Science*, 2005,6:256-263.

[2] Yun Seong Song and Metin Sitti. STRIDE: A Highly Maneuverable and Non-Tethered Water Strider Robot. 2007 IEEE International Conference on Robotics and Automation. Roma, Italy, April 2007,10-14.

[3] http://www.mech.chuo-u.ac.jp/~nakalab/study/bm/waterstrider.html.

[4] Girard.Design review #1: water strider. 2005.2. http://www.columbia.edu/~amh2003/

[5] Qinmin Pan,Min Wang. Miniature Boats with Striking Loading Capacity Fabricated from Superhydrophobic Copper Meshes. ACS Applied Materials & Interfaces. 2009, 1 (2), 420-423.

[6] Bush, John W.M. David L. Hu. Brian Chan. The hydrodynamics of water strider locomotion. Nature, 2003(424): 663–666.

[7] X.Gao, L. Jiang. Water-repellent legs of water striders. *Nature*, 2004(432), 36.

[8] Gao Tiehong, Cao Junyi, Zhu Dunyu. Study on Kinematics analysis and mechanism realization of a novel robot walking on water surface. IEEE-ICIT2007, pp685-690.

[9] Gao Tiehong. Analysis and research on the water stride robot mechanism and properties. Dissertation of Doctor Degree of Hebei University of Technology.2008.

[10] Gao Tiehong, Cao Junyi, Gao Feng, et al. The research of a bionic robot that can walk on watersurface based on water strider. *The International Technology and Innovation Conference 2006*.Hangzhou,China, IET 2006: 214.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 443-447

Three Dimensional Self Calibration Guidance Law for Guided Munitions

Shengqi Chen , and Jun Zhou
College of Astronautics/ Northwestern Polytechnical University , Xi'an,China
E-mail: xigongda_dubo@163.com

Abstract—**In this paper, one kind of three dimensional guidance law for the air-to–surface attacking is derived based on the variable structure control (VSC) theory, then the relationship between the measurement error and guidance error is investigated and discussed thoroughly, the 3-D guidance law based on VSC is modified in order to be implemented simply，the performance of modified guidance law is validated through the comparison between traditional guidance law and modified guidance law based on the computer simulation.**

Index Terms—**Measurement error, Guidance error, 3-Dguidance law, Variable structure control**

I. INTRODUCTION

As to the guidance law of various flying vehicles, many guidance laws have been proposed and applied in the research papers and practical engineering, such as guidance laws based on variable structure control [1~2], optimal control laws [3], differential game laws [4] and known classical guidance laws [6]. All these guidance laws were applied in two dimensional cases and three dimensional scenarios. In order to enhance the efficient of munitions, the development of guidance law research has three typical features: one is more and more constraints were considered in the design of the guidance laws, such as impact angle and multiple target attack [6~7]. The other is modified guidance laws and integrated guidance laws were proposed and simulated, for example, predictive guidance laws based on proportional navigation law [8] .The last one is more and more guidance laws become less depending on the flight time and other uncertainties. Although some uncertainties were taken into consideration during the design period, the disturbances and uncertainties were considered as one undependable term in the formulation, actually these disturbances and uncertainties were related and dependable theoretically and practically, because in the practical application or engineering, various measurement errors are inevitable and these errors have effects on the guidance performance, especially the miss distance. How measurement errors affect the guidance error qualitatively and quantitatively is the main purpose of this paper. It should be noted that in this paper, the measurement errors include the position errors and velocity errors of both munitions and targets. Also in this paper, the guidance errors include the distance error, the light of sight angle errors and rat of light of sight angle errors between the munitions and target.

This paper is organized as follows: the three dimensional guidance law was presented in the section Ⅱ, error calibration was researched in the section Ⅲ, modified guidance law was proposed in the section Ⅳ, simulation was performed in the section Ⅴ, and the conclusion was given finally.

II. THREE DIMENSIONAL GUIDANCE LAW

Suppose that (x_i, y_i, z_i) is the nominal position of the munitions in the local level frame, (x_m, y_m, z_m) is the measured position of munitions by the strapdown inertial navigation system boarded on the munitions. (x_T, y_T, z_T) is the target position in the local level frame. From these positions, the nominal distance and related LOS angle can be depicted as

$$\begin{cases} R_i = \sqrt{(x_T - x_i)^2 + (y_T - y_i)^2 + (z_T - z_i)^2} \\ q_{\varepsilon i} = arctg[\dfrac{(y_T - y_i)}{\sqrt{(x_T - x_i)^2 + (z_T - z_i)^2}}] \\ q_{\beta i} = arctg[\dfrac{z_i - z_T}{x_T - x_i}] \end{cases} \quad (1)$$

From the actual measured data, the measured distance and related LOS angle will be of following forms:

$$\begin{cases} R_m = \sqrt{(x_T - x_m)^2 + (y_T - y_m)^2 + (z_T - z_m)^2} \\ q_{\varepsilon m} = arctg[\dfrac{(y_T - y_m)}{\sqrt{(x_T - x_m)^2 + (z_T - z_m)^2}}] \\ q_{\beta m} = arctg[\dfrac{z_m - z_T}{x_T - x_m}] \end{cases} \quad (2)$$

In the following discussion, all the variables with subscription i means the nominal values, all the variables with subscription m means the measured values. The nominal value and measured value meet the following constraints:

$$\begin{aligned} x_i &= x_m - \Delta x \quad y_i = y_m - \Delta y \\ z_i &= z_m - \Delta z \quad R_i = R_m - \Delta R \end{aligned} \quad (3)$$

Chen Shengqi ,1965, His current research interests include: controlling and guidance of the spacecraft.

AP-PROC-CS-09CN005

Δx, Δy, Δz, ΔR denote the position errors of munitions and relative distance error between munitions and target.

In the practical scenario, the posit parameters of target have been imbedded in the computer boarded on the munitions, so we can obtain following equations based on equation (1):

$$\begin{cases} \dot{R}_i = -\dfrac{(x_{Ti}-x_i)}{R_i}\dot{x}_i - \dfrac{(y_{Ti}-y_i)}{R_i}\dot{y}_i - \dfrac{(z_{Ti}-z_i)}{R_i}\dot{z}_i \\ \quad = -\dot{x}_i\cos q_{\varepsilon i}\cos q_{\beta i} - \dot{y}_i \sin q_{\varepsilon i} + \dot{z}_i \cos q_{\varepsilon i}\sin q_{\beta i} \\ \dot{q}_{\varepsilon i} = \dfrac{\dot{x}_i \sin q_{\varepsilon i}\cos q_{\beta i} - \dot{y}_i \cos q_{\varepsilon i} - \dot{z}_i \sin q_{\varepsilon i}\sin q_{\beta i}}{R_i} \\ \dot{q}_{\beta i} = \dfrac{\dot{x}_i \sin q_{\beta i} + \dot{z}_i \cos q_{\beta i}}{R_i \cos q_{\varepsilon i}} \end{cases} \tag{4}$$

Differentiating equation (4), we can obtain:

$$\begin{cases} \ddot{R}_i = R_i\dot{q}_{\varepsilon i}^2 + R_i\dot{q}_{\beta i}^2\cos^2 q_{\varepsilon i} - \\ \quad (\ddot{x}_i\cos q_{\varepsilon i}\cos q_{\beta i} + \ddot{y}_i \sin q_{\varepsilon i} - \ddot{z}_i \cos q_{\varepsilon i}\sin q_{\beta i}) \\ \ddot{q}_{\varepsilon i} = -\dfrac{2\dot{R}_i\dot{q}_{\varepsilon i}}{R_i} - \dot{q}_{\beta i}^2\cos q_{\varepsilon i}\sin q_{\varepsilon i} + \\ \quad \dfrac{(\ddot{x}_i \sin q_{\varepsilon i}\cos q_{\beta i} - \ddot{y}_i \cos q_{\varepsilon i} - \ddot{z}_i \sin q_{\varepsilon i}\sin q_{\beta i})}{R_i} \\ \ddot{q}_{\beta i} = -\dfrac{2\dot{R}_i\dot{q}_{\beta i}}{R_i} + 2\dot{q}_{\varepsilon i}\dot{q}_{\beta i}tgq_{\varepsilon i} + \dfrac{(\ddot{x}_i \sin q_{\beta i} + \ddot{z}_i \cos q_{\beta i})}{R_i \cos q_{\varepsilon i}} \end{cases} \tag{5}$$

Actually in the racket of equation(5), these terms denote the acceleration projection of munitions in the vision frame:

$$\begin{cases} a_{x'} = \ddot{x}_i\cos q_{\varepsilon i}\cos q_{\beta i} + \ddot{y}_i \sin q_{\varepsilon i} - \ddot{z}_i \cos q_{\varepsilon i}\sin q_{\beta i} \\ a_{y'} = \ddot{x}_i \sin q_{\varepsilon i}\cos q_{\beta i} - \ddot{y}_i \cos q_{\varepsilon i} - \ddot{z}_i \sin q_{\varepsilon i}\sin q_{\beta i} \\ a_{z'} = \ddot{x}_i \sin q_{\beta i} + \ddot{z}_i \cos q_{\beta i} \end{cases} \tag{6}$$

Combining equation (5) and (6), we get:

$$\begin{cases} \ddot{R}_i = R_i\dot{q}_{\varepsilon i}^2 + R_i\dot{q}_{\beta i}^2\cos^2 q_{\varepsilon i} - a_{x'} \\ \ddot{q}_{\varepsilon i} = -\dfrac{2\dot{R}_i\dot{q}_{\varepsilon i}}{R_i} - \dot{q}_{\beta i}^2\cos q_{\varepsilon i}\sin q_{\varepsilon i} - \dfrac{a_{y'}}{R_i} \\ \ddot{q}_{\beta i} = -\dfrac{2\dot{R}_i\dot{q}_{\beta i}}{R_i} + 2\dot{q}_{\varepsilon i}\dot{q}_{\beta i}tgq_{\varepsilon i} + \dfrac{a_{z'}}{R_i\cos q_{\varepsilon i}} \end{cases} \tag{7}$$

in the tracking period, munitions must meet $\dot{R}_r \prec 0$.

Let suppose that

$$x_r = [q_{\varepsilon i} \quad q_{\beta i} \quad \dot{q}_{\varepsilon i} \quad \dot{q}_{\beta i}]^T \quad u_i = [a_{y'}, a_{z'}]^T \tag{8}$$

Then the equations discussed above can be depicted as

$$\begin{aligned} \dot{x}_r &= \begin{bmatrix} x_{r3} \\ x_{r4} \\ -2\dot{R}_i x_{r3}/R_i - \dot{x}_{r4}^2\cos x_{r1}\sin x_{r1} \\ -2\dot{R}_i x_{r4}/R_i + 2x_{r3}x_{r4}tgx_{r1} \end{bmatrix} + \\ &\quad \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ -1/R_i & 0 \\ 0 & 1/(R_i\cos x_{r1}) \end{bmatrix} u_i \\ &= f(x_r) + g(x_r)u_i \end{aligned} \tag{9}$$

As can be seen from the equation (9), this formulation is based on the nominal parameters, actually various measurement errors occurred in the measurement procedure, practical guidance law should be deduced from the actual measurement data. Let suppose

$$\begin{cases} x_M = [q_{\varepsilon m}, q_{\beta m}, \dot{q}_{\varepsilon m}, \dot{q}_{\beta m}]^T \\ \Delta x_r = x_M - x_r \end{cases} \tag{10}$$

Related guidance law equation can be described as

$$\dot{x}_M = f(x_M) + g(x_M)u_M \tag{11}$$

Compared with equation (11), some references deduce guidance law from following equation

$$\begin{aligned} \dot{x}_r &= f(x_r) - \Delta\dot{x}_r + \frac{\partial f}{\partial x}\Delta x_r + [g(x_r) + \frac{\partial g}{\partial x}\Delta x_r](u_r + \Delta u) \\ &= [f(x_r) + \Delta f] + [g(x_r) + \Delta g](u_r + w) \\ \Delta f &= \frac{\partial f}{\partial x}\Delta x_r - \Delta\dot{x}_r \quad \Delta g = \frac{\partial g}{\partial x}\Delta x_r \quad w = [g(x_r) + \Delta g]\Delta u \\ \Delta x_r &= [\Delta q_\varepsilon, \Delta q_\beta, \Delta\dot{q}_\varepsilon, \Delta\dot{q}_\beta]^T \\ \Delta\dot{x}_r &= [\Delta\dot{q}_\varepsilon, \Delta\dot{q}_\beta, \Delta\ddot{q}_\varepsilon, \Delta\ddot{q}_\beta]^T \end{aligned} \tag{12}$$

in the equation (12), many high order terms were omitted, the detailed form of $f(x_r)$, Δf and Δg will be discussed below.

$$f(x_r) = \begin{bmatrix} x_{r3} \\ x_{r4} \\ -2\dot{R}_i x_{r3}/R_i - \dot{x}_{r4}^2\cos x_{r1}\sin x_{r1} \\ -2\dot{R}_i x_{r4}/R_i + 2x_{r3}x_{r4}tgx_{r1} \end{bmatrix} \tag{13}$$

$$\Delta f(x_r, \Delta\dot{x}_r) = \frac{\partial f}{\partial x_r}\Delta x_r - \Delta\dot{x}_r$$

$$= \begin{bmatrix} 0 \\ 0 \\ (-\Delta\ddot{q}_\varepsilon - 2\dot{q}_{\beta i}^2\Delta q_\varepsilon\cos 2q_{\varepsilon i} - 2\dot{R}_i\Delta\dot{q}_\varepsilon/R_i) \\ (-\Delta\ddot{q}_\beta - 2\dot{R}_i\Delta\dot{q}_\beta/R_i - 2\dot{q}_{\varepsilon i}tgq_{\varepsilon i}\Delta\dot{q}_\beta \\ +2\dot{q}_{\beta i}tgq_{\varepsilon i}\Delta\dot{q}_\varepsilon + 2\dot{q}_{\varepsilon i}\dot{q}_{\beta i}\sec^2 q_{\varepsilon i}) \end{bmatrix} \tag{14}$$

$$\Delta g=\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \dfrac{\Delta R}{R_i(R_i+\Delta R)} & 0 \\ 0 & \dfrac{\Delta q_\varepsilon \sin q_{\varepsilon i}}{R_i\cos^2 q_{\varepsilon i}}-\dfrac{\Delta R}{R_i(R_i+\Delta R)\cos^2 q_{\varepsilon i}} \end{bmatrix} \tag{15}$$

If we suppose $\eta=\Delta R/R_m$, then (15) can be rewritten as

$$\Delta g=\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \dfrac{\eta}{R_m(1-\eta)} & 0 \\ 0 & \dfrac{\Delta q_\varepsilon \sin q_{\varepsilon i}}{R_m(1-\eta)\cos^2 q_{\varepsilon i}}-\dfrac{\eta}{R_m(1-\eta)\cos^2 q_{\varepsilon i}} \end{bmatrix} \tag{16}$$

Combing the variable structure control theory and equation (9), we get

$$\begin{cases} u_{i1}=(k+2)\left|\dot R_i\right|\dot q_{\varepsilon i}-R_i\dot q_{\beta i}^2\cos q_{\varepsilon i}\sin q_{\varepsilon i}+R_i\varepsilon\,\mathrm{sgn}(\dot q_{\varepsilon i}) \\ u_{i2}=-(k+2)\left|\dot R_i\right|\dot q_{\beta i}\cos q_{\varepsilon i}-2R_i\dot q_{\varepsilon i}\dot q_{\beta i}\sin q_{\varepsilon i}- \\ \quad R_i\varepsilon\cos q_{\varepsilon i}\,\mathrm{sgn}(\dot q_{\beta i}) \end{cases} \tag{17}$$

As can be seen from the equation (14) and (16), it's obviously unreasonable to treat the various disturbances and uncertainties as one term in the equation, because the guidance error can be resulted from errors, such as position measurement error, velocity measurement error and acceleration measurement error.

Guidance law deduced from (9) can be seen from many references, in this paper, we propose a guidance law deduced from equation (11).

Suppose the sliding hyper surface is

$$S_m=\begin{bmatrix} 0&0&1&0 \\ 0&0&0&1 \end{bmatrix}x_M=Cx_M=\begin{bmatrix} \dot q_{\varepsilon m} \\ \dot q_{\beta m} \end{bmatrix} \tag{18}$$

According to the VSC theory

$$\dot S_m=-k\frac{\left|\dot R_m\right|}{R_m}S_m-\varepsilon\,\mathrm{sgn}(S_m) \tag{19}$$

Then we get

$$\begin{cases} u_{m1}=(k+2)\left|\dot R_m\right|\dot q_{\varepsilon m}-R_m\dot q_{\beta m}^2\cos q_{\varepsilon m}\sin q_{\varepsilon m} \\ \quad +R_m\varepsilon\,\mathrm{sgn}(\dot q_{\varepsilon m}) \\ u_{m2}=-(k+2)\left|\dot R_m\right|\dot q_{\beta m}\cos q_{\varepsilon m}-2R_m\dot q_{\varepsilon m}\dot q_{\beta m}\sin q_{\varepsilon m} \\ \quad -R_m\varepsilon\cos q_{\varepsilon m}\,\mathrm{sgn}(\dot q_{\beta m}) \end{cases} \tag{20}$$

From (19), we can get practical guidance law from real measurement data.

III. ERRORS CALIBRATION

If we can get some prior knowledge about $\Delta P,\Delta V$, then we can calibrate guidance error theoretically according to figure 1.

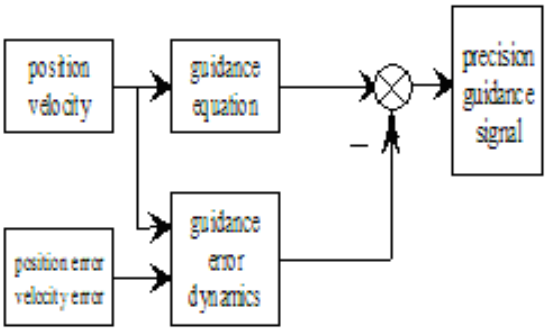


Figure 1. Guidance error calibration scheme

After calibration, the precision control input is

$$\begin{cases} u_{i1}=(k+2)\left|\dot R_m-\Delta\dot R\right|(\dot q_{\varepsilon m}-\Delta\dot q_\varepsilon)- \\ \quad 0.5\times(R_m-\Delta R)(\dot q_{\beta m}-\Delta\dot q_\beta)^2\sin 2(q_{\varepsilon m}-\Delta q_\varepsilon) \\ \quad +(R_m-\Delta R)\varepsilon\,\mathrm{sgn}(\dot q_{\varepsilon m}-\Delta\dot q_\varepsilon) \\ u_{i2}=-(k+2)\left|\dot R_m-\Delta\dot R\right|(\dot q_{\beta m}-\Delta\dot q_\beta)\cos(q_{\varepsilon m}-\Delta q_\varepsilon) \\ \quad -2(R_m-\Delta R)(\dot q_{\varepsilon m}-\Delta\dot q_\varepsilon)(\dot q_{\beta m}-\Delta\dot q_\beta)\sin(q_{\varepsilon m}-\Delta q_\varepsilon) \\ \quad -(R_m-\Delta R)\varepsilon\cos(q_{\varepsilon m}-\Delta q_\varepsilon)\,\mathrm{sgn}(\dot q_{\varepsilon m}-\Delta\dot q_\varepsilon) \end{cases} \tag{21}$$

If we deduce the ideal control input from the actual measurement data, then

$$\begin{aligned} u_{i1}=&\frac{(R_m-\Delta R)}{R_m}u_{m1}+(R_m-\Delta R)\Delta\ddot q_\varepsilon \\ &+2\left(\frac{\dot R_m\dot q_{\varepsilon m}\Delta R}{R_m}-\dot R_m\Delta\dot q_\varepsilon-\dot q_{\varepsilon m}\Delta\dot R+\Delta\dot R\Delta\dot q_\varepsilon\right) \\ &-\dot q_{\beta m}^2\Delta q_\varepsilon\cos 2q_{\varepsilon m}-\dot q_{\beta m}\Delta\dot q_\beta\sin 2q_{\varepsilon m} \\ &-0.5\dot q_{\beta m}^2\Delta q_\varepsilon^2\sin 2q_{\varepsilon m}+2\dot q_{\beta m}\Delta\dot q_\beta\Delta q_\varepsilon\cos 2q_{\varepsilon m} \\ &+0.5\Delta\dot q_\beta^2\sin 2q_{\varepsilon m}+\dot q_{\beta m}\Delta\dot q_\beta\Delta q_\varepsilon^2\sin 2q_{\varepsilon m} \\ &-\Delta\dot q_\beta^2\Delta q_\varepsilon\cos 2q_{\varepsilon m}-0.5\Delta\dot q_\beta^2\Delta q_\varepsilon^2\sin 2q_{\varepsilon m} \end{aligned} \tag{22}$$

$$\begin{aligned} u_{i2}=&\frac{(R_m-\Delta R)}{R_m}\frac{\cos(q_{\varepsilon m}-\Delta q_\varepsilon)}{\cos q_{\varepsilon m}}u_{m2} \\ &-(R_m-\Delta R)\cos(q_{\varepsilon m}-\Delta q_\varepsilon)\Delta\ddot q_\beta \\ &+2\frac{\dot R_m\dot q_{\beta m}\Delta R}{R_m}(\cos q_{\varepsilon m}+\Delta q_\varepsilon\sin q_{\varepsilon m}) \\ &+2(\Delta\dot R\Delta\dot q_\beta-\dot R_m\Delta\dot q_\beta-\dot q_{\beta m}\Delta\dot R)\cos q_{\varepsilon m} \\ &-2(\dot R_m\Delta\dot q_\beta\Delta q_\varepsilon+\dot q_{\beta m}\Delta\dot R-\Delta\dot R\Delta\dot q_\beta)\Delta q_\varepsilon\sin q_{\varepsilon m} \\ &+2(R_m-\Delta R)[(\dot q_{\varepsilon m}\dot q_{\beta m}-\dot q_{\varepsilon m}\Delta\dot q_\beta-\dot q_{\beta m}\Delta\dot q_\varepsilon \\ &+\Delta\dot q_\beta\Delta\dot q_\varepsilon)\Delta q_\varepsilon\cos q_{\varepsilon m}+(\dot q_{\varepsilon m}\Delta\dot q_\beta+\dot q_{\beta m}\Delta\dot q_\varepsilon- \\ &\Delta\dot q_\beta\Delta\dot q_\varepsilon)\sin q_{\varepsilon m}] \end{aligned} \tag{23}$$

From the equation (22) and (23), we can obviously realize the relationship between the guidance error and measurement errors, furthermore, we can understand the effect of measurement error to the guidance law.

IV. MODIFIED GUIDANCE LAW

The control inputs in the equation(17) and (20) are orthogonal accelerations in the vision frame, these parameters should be related with command angular rate of munitions, whereas the angular rate vector is in the body frame, so transition equations should be given below

$$\begin{cases} a_{xb} = V[\dot{\psi}(\sin\beta\cos\vartheta\cos\gamma - \sin\alpha\cos\beta\cos\vartheta\sin\gamma) \\ \quad + \dot{\vartheta}(\sin\alpha\cos\beta\cos\gamma + \sin\beta\sin\gamma)] \\ a_{yb} = V[\dot{\vartheta}\cos\alpha\cos\beta\cos\gamma - \dot{\gamma}\sin\beta \\ \quad - \dot{\psi}(\cos\alpha\cos\beta\cos\vartheta\sin\gamma + \sin\beta\sin\vartheta)] \\ a_{zb} = -V[\dot{\psi}(\cos\alpha\cos\beta\cos\vartheta\cos\gamma + \sin\alpha\sin\vartheta\cos\beta) \\ \quad + \dot{\vartheta}\cos\alpha\cos\beta\sin\gamma + \dot{\gamma}\sin\alpha\cos\beta] \end{cases} \tag{24}$$

Now the angle of attack and angle of sideslip have been considered, the accelerations in the (24) should be transformed to the vision frame

$$a_v = R_L^v R_b^L a_b \tag{25}$$

Equation (25) can be briefly noted as

$$a_v = R_\omega^v \begin{bmatrix} \dot{\vartheta} & \dot{\psi} & \dot{\gamma} \end{bmatrix}^T \tag{26}$$

Where the matrix R_ω^v is a transition matrix, its elements are functions of attitude angle, sideslip angle ,angle of attack and angle of LOS. Its detailed form can be described as

$$R_\omega^v = V \begin{bmatrix} R_{11}'' & R_{12}'' & R_{13}'' \\ R_{21}'' & R_{22}'' & R_{23}'' \\ R_{31}'' & R_{32}'' & R_{33}'' \end{bmatrix} \tag{27}$$

It should be noted that in the equation (27), each element is very complicated, it's not easy to handle in the practical engineering, so these formulations should be modified.

Traditional proportional guidance law can be noted as

$$\dot{\theta}_V = k_{v\varepsilon}\dot{q}_\varepsilon \qquad \dot{\psi}_V = k_{v\beta}\dot{q}_\beta \tag{28}$$

Where $k_{v\varepsilon}$ $k_{v\beta}$ are proportional coefficients and often these parameters are given certain values. Actually the vector of velocity of munitions and LOS is not in the same plane during the flight of munitions. So we modified equation (24) and present following simplified form

$$\begin{cases} a_y = (\dot{\vartheta} - \dot{\alpha})V\cos(q_\varepsilon - \vartheta + \alpha) \\ a_z = -(\dot{\psi} - \dot{\beta})V\cos(q_\beta - \psi + \beta) \end{cases} \tag{29}$$

$$\begin{cases} \dot{\vartheta}_c = \dot{\alpha} + \dfrac{[(k+2)\left|\dot{R}\right|\dot{q}_\varepsilon - R\dot{q}_\beta^2\cos q_\varepsilon \sin q_\varepsilon + R\dot{\varepsilon}\,\mathrm{sgn}(\dot{q}_\varepsilon)]}{V\cos(q_\varepsilon - \vartheta + \alpha)} \\ \dot{\psi} = \dot{\beta} + \dfrac{[(k+2)\left|\dot{R}\right|\dot{q}_\beta + 2R\dot{q}_\varepsilon\dot{q}_\beta \sin q_\varepsilon + R\dot{\varepsilon}\cos q_\varepsilon\,\mathrm{sgn}(\dot{q}_\beta)]}{V\cos(q_\beta - \psi + \beta)} \end{cases} \tag{30}$$

Compared equation (30) with equation (28), we can see several nonlinear modified terms be added to the traditional proportional guidance laws.

V. SIMULATION

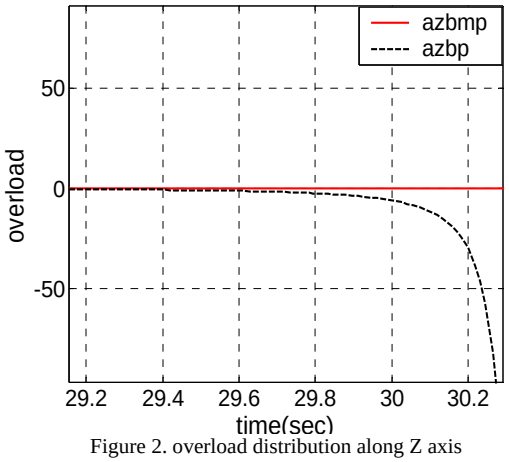


Figure 2. overload distribution along Z axis

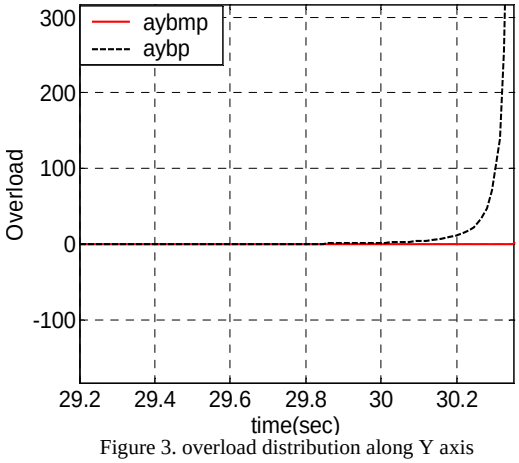


Figure 3. overload distribution along Y axis

It's well known that the quantitative relationship between measurement error and guidance error depends on the flight parameters of munitions, for the brevity, this paper performed the comparison of the overload in the equation (28) and (30) by the simulation under MATLAB/SIMULINK. In the simulation, the chosen munitions is unpowered flying vehicle, related parameters can be found in the reference [9] .the simulation results can be seen from the figure 2 and figure 3. In the figure 2, the red line denotes overload distribution along Z axis of munitions body frame using the modified proportional guidance law, the black line denotes the overload distribution along Z axis of munitions body frame using the traditional proportional

guidance law, it's obviously can be seen that modified guidance law is superior to the traditional guidance in the overload distribution along both Z axis and Y axis, in the above simulation, the proportional coefficients are chosen as 3.

From the results of the simulation, we can see that the maximum overload is 10 using modified guidance law, whereas this value becomes 300 when using traditional guidance law, it should be noted that in the above simulation and results discussion, the given overload value is more command overload than actual necessary overload.

VI. CONCLUSIONS

This paper presents three dimensional self calibration guidance law, it has obvious superiority over traditional guidance laws in the overload distribution and effects of measurement errors and coupled relationship between the flight parameters and the measurement errors. It also has some drawbacks such as more information should be supported by SINS or other information center, this inevitable increase the price of the munitions and complexity of flight control system of munitions.

High precision guided munition is one of the key weapons in the future, so much research topics could be deduced from the result of this paper.

REFERENCES

[1] She wen-xue and Zhou feng-qi, High precision 3D nonlinear variable structure guidance law for homing missile, Journal of astronautics[J]，2004，25（6）：681－685.

[2] Jia qing-zhong and Liu yong-shan, variable structure back stepping guidance law with terminal angular constraint for video-guided penetrating bomb, Journal of astronautics [J]，2008，29（1）：208－214.

[3] Chen ke-jun and Zhao han-yuan, an optimal reentry maneuver guidance law applying to attack the ground target, Journal of astronautics [J], 1994，15(1)：1－8.

[4] Tang shan-tong, the study of load balancing algorithm in on-board computer system based on biased information, Journal of astronautics [J]，2002，23（6）：38－42.

[5] Zarchan, P, Tactical and strategic missiles guidance, AIAA, 1994.

[6] Kim, K.S and Kim, Y, Design of generalized conceptual guidance law using aim angle, Control engineering practice [J], 2004, 12(3): 291－298.

[7] Song, T.L and Shin, S.J, Impact angle control for planar engagement, IEEE transaction on aerospace and electronic systems [J], 1999,35(4):1439－1444

[8] Talole, S.E and Banavar, R.N, Proportional navigation through predictive control, Journal of guidance, control and dynamics [J], 1998,21(6)：1004－1006.

[9] Liu zhi-ping, Zhou feng-qi and Zhou jun, variable structure control for attitude of flying vehicle with parameter uncertainties, Journal of astronautics [J]，2007，28(1).

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 448-452

Design and Optimization of Reentry Trajectory of Maneuverable Warhead

Kaibo Bi[1], Xingbao Yang [2], Zhou Zhou[3], and Chuangang Zhang [4]
[1,2,4] Dalian Naval Academy/ Dept. of Missile and Shipborne Gun, Dalian,China; [3] Dalian Naval Academy/ International military exchanges department, Dalian,China
[1]Email: bkp2004@sina.com

***Abstract*—A strike scheme is presented by restricting the entry velocity , flight path angle and introducing the conception of the best attack line. Based on this scheme, a design method of reentry trajectory is proposed when the earth is assumed to be non-rotating. The method is based on the reentry with maximum left-to-drag, but a small attack angle flight phase and a terminal adjust flight phase are included in. A feasible reentry trajectory with two or three skips can be get by adjust the small attack angle flight time. Based on this trajectory, a standard trajectory considering earth rotation and practical parameters on entry point.**

***Index Terms*—Maneuverable Warhead, strike scheme, reentry trajectory**

I. INTRODUCTION

With the development of the anti-missile system, the existing probability of the ballistic missile is gradually decreased[1] as for its long duration of the booster phase and inertia flight phase. Comparatively speaking, the maneuvering reentry vehicle(MaRV) has its brilliant combat capability. MaRV could be launched as space operation vehicle(SOV) from the near surface trajectory or launched as a small sized carrier missile from ground and guide it to a sub-orbital trajectory. MARV flies into the dense atmosphere by inert, afterwards depending on its own high lift and resistance ratio gliding maneuvering to the designated position and release its military carriers. The anti-surface flight process of MaRV launched from SOV could be seen from Fig. 1.

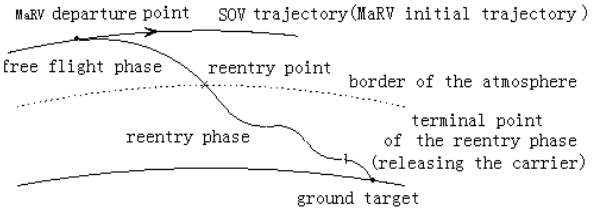


Figure 1. the flight process of anti ground attack by MaRV

II. REENTRY ATTACKING PROGRAM

From the MaRV operation process, one can figure out that the reentry point is the connecting point between the free flight phase and reentry phase, at the departure point applied different departure speed gain Δv in order to obtain different transfer trajectory, corresponding to the reentry point parameter, and affecting the gliding maneuverability at the layers of atmosphere. This article mainly discusses the simplification of the reentry strike scheme by way of limiting the parameter of reentry point.

A. parameter of reentry point on the supposition that the earth is static

The magnitude of the reentry point speed v_e and the track bearing γ_e has great affect to the reentry flight of the MaRV, similiarly, the max available departure speed gain $\Delta v_{\max}$ of MaRV is fix, the value of this two parameter has certain limitations. Reference[4]take the reentry angle within the range of $[-6^\circ \sim -2^\circ]$ while doing the attacking simulation. While reference[5] pointed out that via simulation the best reentry angle should limit to the range of $[-5^\circ, -1.5^\circ]$. This article will tries to determine the reentry point parameter by means of the following simulation methods. Considering the reentry attacking process regardless of the change of the trajectory plane. As per one time impulse departure, applied different braking direction and breaking speed gain at the departure point, different reentry point parameter could be obtained, regarding the specific mathematical expectations of the MaRV , calculate the max lateral maneuvering flight performance corresponding to the inner layer of the atmosphere. Since the max lateral maneuvering range is more at the inner layer of atmosphere, the flight time at the outer layer of the atmosphere is short, the required speed gain of departure is less, a better group of reentry point parameter could be selected from the multi simulation results.

Regarding the different original trajectory of the close ground MaRV, the different reentry point parameter could be determined via simulation methods. This article is on the supposition that the MaRV is moving along the round trajectory at the height of 3000 km, the max seed gain of MaRV is $\Delta v_{\max} = 1000 m/s$, by doing a great amount of departure simulations and experiments of the reentry flying vehicles the selected reentry parameter are $r_e = 6464000m$, $v_e = 7550 m/s$, $\gamma_e = -4.8^\circ$.

B. The reentry attacking planning process based on the optimal attacking line

BI Kaibo,1965, His current research interests include: controlling and guidance of the spacecraft.

On the supposition that the earth is static, once the r_e, v_e and γ_e are determined, the approximately lateral max maneuvering range $\varphi_{1\max}$ of the MaRv at the inner layer of the atmosphere and the corresponding longitudinal flight distance $\theta_{1\max}$ can be obtained by using the max lift and resistance ratio, further more , the flight distance θ_{ex} from the departure point and to the reentry point (expressed by the geocentric angle), flight time t_{ex} and the max variation angle $\beta_{\max}$ of the trajectory plane at the outer layer of the atmosphere are becoming the rated values. The overall lateral max maneuvering range $\varphi_{\max}$ of the warhead section and its corresponding longitudinal range $\theta_{\max}$ from the departure point can be known by taking the use of the numerical value considering the maneuverability at both inner and outer layer of the atmosphere.

Taking the advantage of the lateral and longitudinal maneuvering performance of MaRV after entering into the atmosphere from a certain point, the targets within a large range could be attacked. As the Fig 2 shows: taken the lateral maneuvering range at the inner layer of atmosphere as the radius to make a circle, then there will be great maneuverability for MaRV to attack the targets within the circle at a longer distance, so usually this circle is named as the optimal attacking circle, the center point of this circle is the optimal attacking point. While carrying out the ground target attacking, the targets should be closer to the center area of this circle. This is the starting point of the reentry attacking process. Generally, $\beta_{\max}$ is less, optimal attacking point of same departure point and different reentry point is approximately at one straight line, it is named as the optimal attacking line.

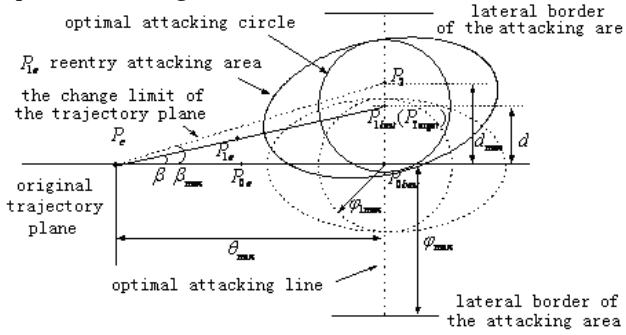


Figure 2. the planned figure of the flight trajectory of MaRV attacking the ground target.

After guided to the concept of "optimal attacking line" taken the time while target is passing through the optimal line as the departure attacking time of MaRV. the intercept range d of the target at the optimal attacking line is determined,(shown in Fig.2) supposed that the intercept range $d_{\max}$ at the attacking line of the trajectory under the dotted the lines after the trajectory plane of MaRV change $\beta_{\max}$ angle, if the $\beta_{\max}$ is less, the variation angle of the required trajectory plane at the departure point of MaRV is :

$$\beta = \begin{cases} \dfrac{d}{d_{\max}}\beta_{\max} & d < d_{\max} \\ \beta_{\max} & d \geq d_{\max} \end{cases} \qquad (1)$$

The conclusion of to -the -ground attacking planned process on the supposition that the earth is static is as follows:

calculate $d_{\max}$, determine the intercept range d and time τ when the target point passing through the optimal attacking line.

Determine position P_c at the departure point according to τ, according to (1) formula determine β so as to obtain the reentry point position P_e.

Select the proper guidance method at the departure section, guide the MaRV entering into the free flight trajectory.

Based on the reentry point parameter and the terminal guidance initial parameter design the trajectory of the reentry section.

If the earth is rotating, the optimal attacking line is perpendicular to the initial trajectory of the dotted line trace of MaRV, neglect the deviations value $d_{\max}$, $\varphi_{\max}$ result from the earth rotation, and convert the reentry point parameter into the earth coordinate, then the above mentioned planned process can still be in use.

III. THE DESIGN OF THE STANDARD TRAJECTORY AT THE REENTRY SECTION

The main objective of the standard trajectory design at the reentry section is to get the trajectory of the reentry section which meets the requirement of the reentry terminal restrictions. under restriction of the aerodynamic heating, pressure, overload. By means of simulation, the MaRV at the inner layer of the atmosphere, according to the max lift and resistance ratio flight obtains the max lateral maneuvering trajectory which meets the requirement of the reentry flight restrictions.

A. terminal parameter of the reentry section

The terminal parameter of the MaRV reentry section relates with the releasing conditions, mainly includes height, speed, track bearing, course angle and missile-target- range. this article directly selects the following terminal status parameter: height $h_f = 25000m$,track bearing $\gamma_f = 0^\circ$,speed $v_f = 1500m/s$,course angel pointed at the target point, missile and target range is $L_f = 100km$.

B. the design of the reentry trajectory on the supposition that the earth is static

Reentry speed of MaRV is close to the approaching trajectory speed, and the speed to release the carrier is comparatively slow, so it is required to consume lots of

initial energy of MaRV to the atmosphere at the reentry section. In a respect of energy, the essence of the reentry trajectory design is the control energy consumption under the condition of a certain restrictions.

In order to have a visual understanding about the max lift and resistance ratio, Fig.3 shows the comparison between the MaRV plane reentry and the max lateral maneuvering reentry. The Fig. illustrates : the first wave trough of the plane reentry should be higher than the max lateral maneuvering reentry. the later has the rolling angle which result in the decrease of the lift component on the vertical plane.

Based on the above mentioned simulating result, this article tries to explain by reducing the flight attacking angle when MaRV first reaches the wave trough at a certain time to prolong the flight time of MaRV at the dense atmosphere so as to consume the mechanical energy effectively. To be conclude, to control the consumption of the energy of MaRV by adjusting the small attacking angel flight time. Suppose that the small attacking angle initial fight time $t_{as} = 75s$, the small attacking angle is $\tilde{a} = 6^o$, Fig. 4 is the reentry comparison of about the sustaining time of the small attacking angle $t_{af} = 80s$, $t_{af} = 100s$ and $t_{af} = 120s$.

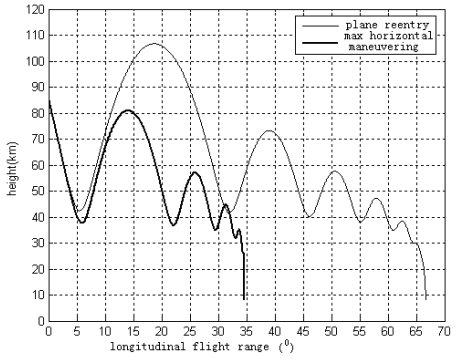


Figure3a. height comparison of max lift and resistance ratio

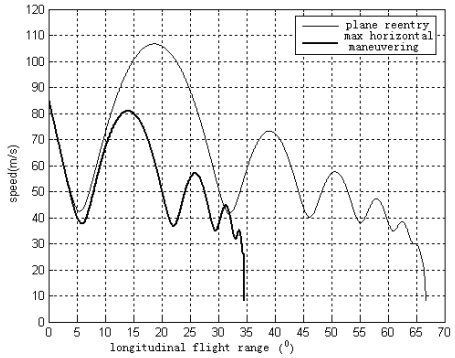


Figure3b. speed comparison of the max lift and resistance ratio

The multi simulation shows it is effective to consume the mechanical energy by reduce the flight time of the attacking angle while reaching the wave trough. The thermal power, pressure, overload peak and overall thermal energy of the reentry process are close to the

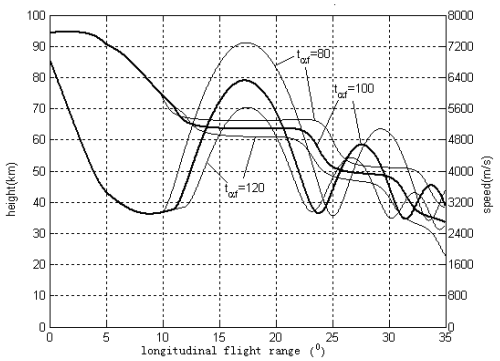


Figure 4. the reentry comparison of different small attacking angle sustaining time.

max lateral maneuvering reentry value, furthermore, $t_{as}, \tilde{a}, t_{af}$ these three parameter could be used to adjust the energy consumption, among of which t_{as} has the most affect on the energy consumption., t_{af} is second. If t_{as} is not properly selected, there is no way to adjust the energy consumptions effectively to another to parameter. So , usually, $\tilde{a}$ is a constant, and while MaRV is making lateral maneuvering reentry adjust t_{as} :

$$t_{\alpha s} = t_{\alpha sp} + d'(t_{l0} - t_{\alpha sp}) / \varphi_{\max} \qquad (2)$$

In this fomular t_{asp} is the initial moment of the small attacking angle when making the plane reentry, t_{l0} is the max lateral maneuvering reentry flight time.

$$d' = \begin{cases} 0 & d \le d_{\max} \\ d - d_{\max} & d > d_{\max} \end{cases} \qquad (3)$$

The terminal speed of the reentry section should be in align with the target. That is the terminal rolling angle is $\sigma = 0^0$. Based on this restrictions, to carry out the corrections of max lateral maneuvering rolling angle, is to make MaRV as zero when it is at the L_f the optimal attacking line. MaRV ‘s longitudinal flight distance are all $\theta_{1\max}$ within the atmosphere, so the variation law of the rolling angle could be expressed by the longitudinal flight range at the inner layers of atmosphere. when MaRV is the max reentry lift and resistance ratio, the max lateral maneuvering rolling angle could be expressed in this way.

$$\sigma_{c\max}(\theta) = \begin{cases} \sigma_{c\max 0}(\theta) & \theta < \theta_z \\ 0 & \theta \ge \theta_z \end{cases} \qquad (4)$$

Among of which $\theta_z = \theta_{1\max} - 57.3 L_f / R_E$

Rolling angel is the one which corresponding to the max lateral maneuvering in a fix proportion, that is :

$$\sigma(\theta) = p\sigma_{c\max}(\theta) \qquad (5)$$

In this formula, p is the proportional coefficient, $p \in [0,1]$. Once the variation law of the attacking angle is determined, regarding the target at the intercept from the

optimal attacking line, by using the mathematical calculation the corresponding p value could be obtained in order to guide the MaRV dotted trajectory passing through the target point. So, once the target at the intercept from the optimal attacking line, only by adjusting the attacking flight time t_{af} , the energy consumption could be controlled . the flight trajectory of MaRV which was passed by the dotted trajectory after adjusting the small attacking angle is named as the small attacking angle flight trajectory.

Considering the reentry restriction of the terminal height , speed, course, and on the basis of the small attacking angle flight trajectory, the terminal adjusting flight section is discussed here. the ground projection of the terminal adjusting flight section is the dotted trace of the small attacking angle flight trajectory. To be sure that the flight status parameter has the secondary continuity, the projection curve on the longitudinal flight plane of this trajectory is the third curve, at the deduction section of the third and fourth flight trajectory finds out the starting point of terminal adjusting phase. Take the geocentric range as The cubic polynomial of the longitudinal flight distance:

$$r(\theta) = a_3\theta^3 + a_2\theta^2 + a_1\theta + a_0 \qquad (6)$$

The equation coefficient can be determined by the following conditions:

$$\begin{cases} 3\theta_0^2 a_3 + 2\theta_0 a_2 + a_1 = r_0' \\ 6\theta_0 a_3 + 2a_2 = r_0'' \\ \theta_0^3 a_3 + \theta_0^2 a_2 + \theta_0 a_1 + a_0 = r_0 \\ 3\theta_1^2 a_3 + 2\theta_1 a_2 + a_1 = r_1' \end{cases} \qquad (7)$$

In this equation θ_0 、 θ_1 is the starting point and terminal point of the terminal adjusting phase(starting point positions at the deduction section of the third and forth trajectory, terminal point is the starting point of the terminal guidance) r_0' 、 r_0'' is the first and second order derivative of the longitudinal range corresponding to the geocentric range at the starting point, by MaRV movement parameter calculation could be obtains, r_1' is the first order derivative of the longitudinal range corresponding to the geocentric range at the terminal point under the terminal restrictions.

So far the non linear relations between the small attacking angle flight time and terminal speed have established. According to the several times adjustment of the small attacking angle flight time while the target at the optimal attacking line , could determine, the starting point at the terminal adjusting phase is at the third deduction section or the forth, similarly the small attacking angle flight time interval which includes the terminal restricted speed is determined. What is more, as this design phase regardless the rotation of the earth, so the various flight time interval could be tested at the different intercept. Since this is known, by using the non linear equation the small attacking angle flight time which terminal restricted speed is 1500m/s could be searched, the reentry trajectory with terminal restrictions can be obtained.

C. the standard trajectory design on the supposition that the earth is rotating.

Supposed that the reentry trajectory is trj0 under the condition of the earth is static, the control quantity of trajectory $trj0$ (angle of attack, roll angle) within time area parameterized. And convert the optimal control into the parameter optimization, the time range is more than the actual control time, the small angle flight section is the important one, the parameter optimization after conversion could be expressed in the following way:

$$\begin{aligned} &\min f(\mathbf{x}) \qquad \mathbf{x} \in [\mathbf{a}, \mathbf{b}] \\ &s.t. \qquad \mathbf{g}(\mathbf{x}) - \mathbf{g}_{\max} \le 0,\ \mathbf{h}(\mathbf{x}) - \mathbf{h}_f = 0 \end{aligned} \qquad (8)$$

Among of which $x = [a_1, \cdots, a_N, \sigma_1, \cdots, \sigma_N]^T$ as the parameter variations, $a_{\min} \le a_i \le a_{\max}$, N 、 $a_{\min}$ 、 $a_{\max}$ 、 $\sigma_{\min}$ 、 $\sigma_{\max}$ the dispersion point within a certain time range and the limit of attacking angle and roll angle respectively, is $g_{j\max}$ 、 $g_i(x)$ the JTH restricted value and the actual value of X in the course of the reentry. h_{jf} 、 $h_j(x)$ is the KTH terminal restricted value and actual value $\mathbf{x}$ at the reentry process if the optimal parameter is x .

Based on the strict restrictions of the optimization, the initial parameter which meet the restriction requirement is difficult to determine, only take the non-feasible region as the optimal initial value, the feasible region which meet the restriction requirement is very limit, and within this limit feasible region it is difficult to increase the performance index. So the main objective of this article is : starting from some non –feasible initial value, tries to find out the feasible solution which meet the restriction requirement, while taking the optimal index as the min reentry flight time.

Take the above parameter as the initial value, considering the earth is rotating, reentry restrictions and terminal restrictions, adopting the parameter optimization solution (8) to get the reentry stardard trajectory trj^0 . But the traditional optimal method is sensitive to the initial value, After the control quantity of trajectory $trj0$ is changed into parameters, optimization trajectory trj^1 is obtained on the supposition that the earth is static, then take this value as the initial value, then convert this solution into the rotating conditions obtain trj^0 .

IV. CALCULATION OF THE REENTRY ATTACK PLAN

Supposed that the MaRV is moving on the round trajectory with the altitude of 300km, trajectory deflection angle is 55^0, right ascension of the ascending node(RAAN) 0^0 is. the argument of perigee is 0^0 . the true anomaly at the initial moment

is 0^0 . $\varphi_{1\max}=14^0, \theta_{1\max}=35^0$ the longitudinal range between the optimal attacking line and MaRV is $\theta^*=70.17^0$. The ground target longitude is $\theta_T=165^0$, latitude is $\varphi_T=35^0$.

TABLEI. PARAMETERS ON THE PROCESS OF REENTRY ATTACKING PLAN

$\beta_{\max}$	$d_{\max}$	d	f_c	$T_{orbit}(s)$	β	$t_{ex}(s)$
6.7793°	6.3807°	12.53°	75.10°	227.8	6.7793°	548.7

$v_e(m/s)$	γ_e	ψ_e	r_e	θ_e	φ_e
7294.8	−4.969°	−23.71°	6464000	122.17°	53.76°

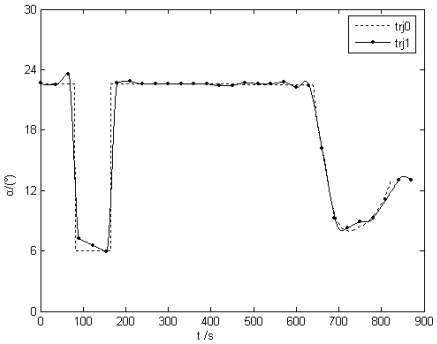


Figure5a. comparison of the trajectory of trj0 and trj1 (angle of attack)

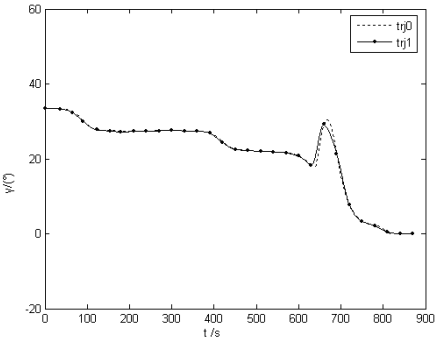


Figure5b.comparison of the trajectory of trj0 and trj1 (roll angle) .

Based on above mentioned plan, part of the parameter will be listed in the table 1, here f_c is the true anomaly of MaRV departure point at the original trajectory, T_{orbit} is the moving time of the MaRV at the original trajectory from the initial time till the attacking time $[v_e, \gamma_e, \psi_e, r_e, \theta_e, \varphi_e]$ are the reentry point parameter in the earth coordinate. Reentry trajectory design takes $t_{asp}=75s$ 、 $t_{l0}=90s$ 、 $\tilde{a}=6^\circ$ while standard trajectory design adopts the secondary solution. Fig5a and Fig 5b is the comparison of the parameter variations on the supposition that the earth is static, Fig 6 is final result of the reentry standard trajectory

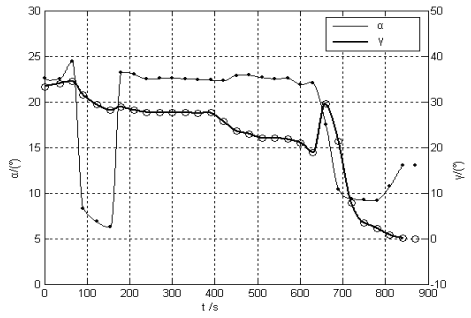


Figure6a. the control quantity of trajectory trj° .

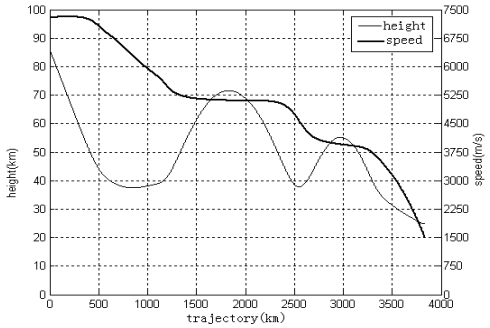


Figure6b. the altitude and speed variations of trajectory tri°

V.CONCLUSION

This article mainly discusses about the operational plan of the MaRV, the specific attacking plan and scheme on the ground target has been addressed here, based on the attacking scheme, the author proposed a certain design method regarding the reentry standard trajectory. By using the specific calculations the accuracy at the reentry terminal section increases.

REFERENCES

[1] GU xian-liang, GONG chun-hua, WU wu-hua. Design and Optimization of Wavy Trajectory for Ballistic Missiles[J].Journal of china ordnance,2005,26（3）:353-356.

[2] George Richie. The common aero vehicle: space delivery system of the future[A].AIAA Space Technology Conference and Exposition[C].Albuquerque,28-30 Sept.1999,99-4435.

[3] Phillip E P. Component based simulation of the space operations vehicle and the common aero vehicle[D]. Master Thesis, United states navy B A, San Diego state University,1999.

[4] ZHANG yan, FENG shu-xing, HE zhong-long. Dynamics Analysis and Modeling Simulation for the Flight Course of the Space-based Strike Weapon. Aerospace control,2006,24(1)：52-56.

[5] CHEN hong-bo,YANG Di. De-Orbit Operation Study of Lifting Reentry Vehicle[J]. Flight Dynamics, 2006,24(2):35-39

[6] RUAN chun-rong. Optimal trajectories in atmospheric flight[M. Beijing: China Astronautic Publishing House,1987.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 453-457

Application of Mutations Progression Method in Enterprise Human Resources Evaluation

Jinying Li, and Zhike Zhang
North China Electric Power University (Baoding), Baoding, China
E-mail：zhangzhike2007@126.com , zzk3611@126.com

***Abstract*—As a study of the mutations phenomenon of system theory,the theory of mutations is widely used in many subjects. One common applications, is derived by mutations model progression method to solve the multi-criteria decision problem. The main characteristics of mutations progession method is that it firstly for the evaluation target system of multi-level contradiction decomposition, and the use of mutations theory and fuzzy methods generates the fuzzy membership function mutation, by concentratione formula for comprehensive quantitative operations, finally as a parameter, namely the general membership functions, which are evaluated. In this paper, Mutations progression method on the use of Enterprise Human Resources Evaluation has been studied and found the result is more reasonable, objective and feasible for the settlement of the issue of Enterprise Human Resources Evaluation, which provides a new way of thinking.**

***Index Terms*—mutations progression method, human Resources, comprehensive evaluation**

Ⅰ. INTRODUCTION

Human resource is one of the most valuable resources in the world as well as a strategic resource in the 21st century, when performance appraisal plays an increasingly important in human resource management [1]. At present, many scholars to the enterprise human resources performance evaluation are studied. The application of various algorithms constructs the effective model, and made many discoveries. Zhijun Han and Xiaojun Cai [2] evaluates the quantitative validity of performance appraisal in human resources management by the method of analysis of variance, and some canonical results are made that may occur in performance appraisal system. Yanling Xiao, Xiaojing Liu and Jianbo Liu [3] use entropy method to adjust performance indicator weight given by the process, so it lives up to dynamic weight index and gives an example to show how to use entropy method. Jian Li [4] uses the method of ANP to appraise the human resource from the aspects of mentality, competence and knowledge and gets the results that the appraisal will provide a justice to the development of the organizations and the individuals. Xisong Liu, Chunrong Du and Yaowu Wang, et al [5] who base on the human resource management of large-scale projects establish an evaluation system according to the achievement, capacity and manner of evaluated object and prove that the evaluation system and evaluation model are successfully applied in a practical project.

These above-mentioned literatures study the performance evaluation of human resources, and apply different algorithms to evaluate the performance of human resources and establish the model. These documents for performance evaluation of enterprise human resources have made important contribution to the research, and practice guidance is obvious. Each method has its own characteristics, and some methods have problem in determining the right weight problem correctly, and some methods' calculation are more complicated, but there is not a mutations progression method applied to the enterprise human resources evaluation. This method is not used to the weight of indicators, but it takes into account the relative importance of evaluation, thereby reducing the subjectivity without losing scientific, reasonable, simple and accurate calculation of its extensive scope, worth exploring.

Ⅱ. THE BASIC THOUGHTS AND PROCEDURES OF THE MUTATIONS PROGRESSION METHOD

A. Use mutations in the evaluation index system of organizations

Firstly, according to the evaluation purpose, decompose the total evaluation index into multi-level, arranged in a handstand tree target by multi-layer structure, from the total evaluation of the lower targets to the indicators, and gradually into the lower sub-indicators. The original data only need to know the bottom sub-indicators. An index of decomposition was to get more specific index, so, when the quantified decomposed into sub-indicators that can be measured, decomposition can stop. Mutation because a state variable coefficients of control variables is not more than four, so all levels corresponding to the general indicators (a subset of indicators for a single indicator) not more than 4 decompositions.

B. Determine the evaluation index system of mutations in the system type mutations

A total of seven types of mutations in the system, the most common are 3 kinds, which is sharp point mutation system; swallow tail mutation system; butterfly mutation system. Their mathematical models are [6]:

(1) Sharp point mutation system：

$$f(x) = x^4 + ax^2 + bx \tag{1}$$

(2) Swallow tail mutation system：

$$f(x)=\frac{1}{5}x^5+\frac{1}{3}ax^3+\frac{1}{2}bx^2+cx \quad (2)$$

(3) Butterfly mutation system:

$$f(x)=\frac{1}{6}x^6+\frac{1}{4}ax^4+\frac{1}{3}bx^3+\frac{1}{2}cx^2+dx \quad (3)$$

Above $f(x)$ says a system of a state variable x, the potential function of state variable x, coefficient of a, b, c, d said the state variable control variables. The system state variables and the control variables are two aspects of the contradiction.

If an indicator is only broken down into two sub-indicators, the system can be regarded as a sharp point mutation system. If an indicator decomposed into three sub-indicators, the system can be regarded as swallow tail mutation system. If an indicator can be broken down into four sub-indicators, the system can be regarded as butterfly mutation system.

C. Bifurcation equation of mutation system derives concentration formula

Mutation system for the potential function for $f(x)$, according to mutations theory, and its all set point into balance surface, the equation of $f(x)$ for the first derivative of the above-mentioned, that is, the singular point $f'(x)=0$. It's collection of odd through the second derivative of $f(x)$, namely $f''(x)=0$ expunction x, get bifurcation set equation of mutations system. Bifurcation set equation shows that if the control variables satisfy this equation, the system will mutate [7].

Through the decomposition of the form of bifurcation set equations to the concentration formula, the concentration formula will return a system different from the quality control of state variables into the same qualitative state, that is, into a state variable that states the quality.

The bifurcation set equation of sharp point mutation system's decomposition form is [8]:

$$a=-6x^2, b=8x^3 \quad (4)$$

Deducing concentration formula $x_a=a^{1/2}$, $x_b=b^{1/3}$ .In the formula, x_a said that the x value corresponding to a, x_b said that the x value corresponding to b

The bifurcation set equation of swallow tail mutation system's decomposition form is:

$$a=-6x^2, c=-3x^4, \quad b=8x^3 \quad (5)$$

Deducing concentration formula $x_a=a^{1/2}$ $x_b=b^{1/3}$, $x_c=c^{1/4}$ .In the formula, x_a said that the x value corresponding to a, x_b said that the x value corresponding to b, x_c said that the x value corresponding to c.

The bifurcation set equation of butterfly mutation system's decomposition form is:

$$a=-10x^2,\ c=-15x^4,\ b=20x^3,\ d=4x^5 \quad (6)$$

Deducing concentration formula $x_a=a^{1/2}$, $x_b=b^{1/3}$, $x_c=c^{1/4}$, $x_d=d^{1/5}$ .In the formula, x_a said that the x value corresponding to a, x_b said that the x value corresponding to b, x_c said that the x value corresponding to c. x_d said that the x value corresponding to d. Here, concentration formula is a multi-dimensional fuzzy membership function in substance.

D. The use of concentrations formula for comprehensive evaluation

According to multi-objective fuzzy decision theory, on the same program objectives in a variety of circumstances, such as A_1, A_2, $\cdots$, A_m for the fuzzy goals, The ideal strategy is $C=A_1\cap A_2\cap\cdots\cap A_m$, and its membership function is [9]:

$$\mu_x=\mu_{A_1}(x)\wedge\mu_{A_2}(x)\wedge\cdots\wedge\mu_{A_m}(x) \quad (7)$$

Here, $\mu_{A_i}(x)$ is the membership functions of A_i, defined as membership function of this program and membership function which of the minimum target.

For different programs, such as for $G_1, G_2, \cdots, G_n$, Charged G_i membership function $\mu G_i(x)>\mu G_j(x)$, that is said program G_i is superior to the program G_j.

Thus the use of concentration formula under the control of the same object variables (indicators) to calculate the corresponding value should be "large and medium-sized small check" principle. However, for complementary indicators, usually in lieu of using the average in the final comparison object to use of "small in the big check" principle, that is sorting of evaluation objects according to the size of the total score evaluation indexes.

III. The Use Of The Mutations Progression Method For Evaluation The Enterprise Human Resources

A. Establishing Enterprise Human Resources Evaluation System

The investigation of enterprise employees in input index selection should reflect the work condition and consumption of resources. In the working status of staff into the way of questionnaire, the system is mainly based on the quality of employees, considering the work

attitude [10]. In the last ,company human resources evaluation index system according to the requirement of mutations progression method into a multi-level evaluating target structure, and according to the indexes in order of importance, an important index in front row , secondary index in behind(see table 1).

B. Determine Evaluation Target System of All Levels of Mutations in the Type of System

According to the basic principle of mutations progression method, indicators at all levels of mutations in the type of system are given by:

(1) The third grade index system. Efficiency is a swallow tail type for non-complementary; work quality is a sharp point mutations system for complementary; safety accident is a sharp point mutations system for complementary; professional is a sharp point mutations system for complementary; knowledge is a swallow tail type for complementary; express is a sharp point mutations system for complementary; organization is a swallow tail type for non-complementary; team is a sharp point mutations system for complementary; professional is a sharp point mutations system for complementary; initiative is a sharp point mutations system for complementary.

(2) The second grade index system. Performance indexes for non-complementary, swallow tail type; capability indexes for complementary, butterfly mutations system; attitude indexes for non-complementary, swallow tail type.

(3) The top of the enterprise human resources performance evaluation index system for the complementary, swallow tail type.

TABLE Ⅰ

ENTERPRISE HUMAN RESOURCES PERFORMANCE EVALUATION MUTATIONS INDEX SYSTEM AND THE STANDARDIZED DATA

A level Indicator	B level Indicator	C Level Indicator	A	B	C	D	E
performance indicators A_1	efficiency B_1	task completion C_1	1.0000	0.9882	0.9993	0.9981	0.9916
		resource utilization C_2	0.8977	0.9432	0.9810	0.9107	1.0000
		cost C_3	0.7799	0.8345	0.9752	1.0000	0.9612
	work quality B_2	pass rate C_4	0.8751	0.8792	0.8872	1.0000	0.9283
		defect rate C_5	0.1438	0.1190	0.1526	0.0932	1.0000
	safety accidents B_3	accident frequency C_6	0.0000	0.0186	0.0193	0.0114	0.0100
		serious injury rate C_7	0.0000	0.0167	0.0103	0.0208	0.0118
capacity indicators A_2	Professional B_4	business knowledge C_8	1.0000	0.9070	0.8893	0.9810	0.7973
		post skills C_9	0.7021	0.7667	1.0000	0.7231	0.8011
	knowledge B_5	learning ability C_{10}	0.7687	0.7587	0.8102	0.8302	1.0000
		thought C_{11}	0.8573	0.9581	1.0000	0.9321	0.8962
		innovation ability C_{12}	0.9001	0.9328	0.9012	0.8173	1.0000
	ability to express B_6	oral expression C_{13}	0.8317	0.8847	0.8901	1.0000	0.9916
		written expression C_{14}	1.0000	0.9155	0.9409	0.8093	0.9807
	organization B_7	strain capacity C_{15}	0.8402	0.8870	0.5627	1.0000	0.8344
		decision power C_{16}	1.0000	0.7885	0.8370	0.8164	0.9807
		plan ability C_{17}	0.6853	0.8762	1.0000	0.8566	0.7973
attitude A_3	team B_8	obedient spirit C_{18}	0.8877	0.8932	0.9810	1.0000	0.9107
		coordinate spirit C_{19}	0.7021	1.0000	0.9667	0.9231	0.8010
	professional B_9	occupational ethics C_{20}	1.0000	0.9810	0.9930	0.8251	0.8265
		discipline C_{21}	0.9785	0.9623	1.0000	0.9123	0.8992
	initiative B_{10}	attendance rate C_{22}	0.9775	1.0000	0.9810	0.9963	0.9105
		responsibility C_{23}	0.9021	0.9631	0.9521	0.8542	1.0000

C. Use Concentration Formula to Evaluate and Sorting

In order to solve the dimensionless model parameters reunification, each index data needs to do dimensionless processing so that dimension of the data after the elimination of restrictions on the value range of 0 ~ 1.

The better indicator of the type (positive indicators):

$$y_{ij} = \frac{x_{ij}}{\max(x_{ij})} \tag{8}$$

The smaller the better type of indicators (reverse targets):

$$y_{ij} = 1 - \frac{x_{ij}}{\max(x_{ij})} \qquad (9)$$

Of which, $i = 1, 2 \cdots, m$ (m is the index number); $j = 1, 2 \cdots, n$ (n is the index number).

Use concentration formula to calculate mutation series of the control variables indicators of each unit that need evaluation, and then take the mutant series of the subsystem of each unit that is evaluated, and as control variables of the evaluation system indicator on the level.

For example, the quality of the evaluation system, an indicator of control variables out of the mutation type of series as indicators of performance evaluation system of indicators of the quality of the control variables. Series is in accordance with the requirements of mutation progression method and mutation system of non-complementary to "large and medium-sized small check" to check the principle, that is, from a mutation in the smallest class, and mutation system of complementary check the mutant series average. Therefore:

For performance indexes:

$$\begin{cases} X_A = \min(x_a, x_b, x_c) \\ X_B = (x_a + x_b)/2 \\ X_C = (x_a + x_b)/2 \end{cases}$$

For capability indexes:

$$\begin{cases} X_A = (x_a + x_b)/2 \\ X_B = (x_a + x_b + x_c)/2 \\ X_C = (x_a + x_b)/3 \\ X_D = \min(x_a, x_b, x_c) \end{cases}$$

For attitude indexes:

$$\begin{cases} X_A = (x_a + x_b)/2 \\ X_B = (x_a + x_b)/2 \\ X_C = (x_a + x_b)/2 \end{cases}$$

In the first use concentration formula to calculate mutation series of the control variables indicators of each unit that need evaluation, then take mutations series as control variables of top indicators, through mutations series we can get the overall evaluation score of the staffs in the last. This score can evaluate on human resource performance evaluation and can also sort the total score.

D. Enterprise Human Resources Evaluation of Empirical Analysis

For convenience, the author describes a recruitment examination, for example. Assuming a senior management personnel recruitment Of A and B, C, D, E five candidate performance ability, attitude, three kinds of assessment, the application of the method for calculating the standardized series of mutation data in table 1. Each man will only needs to put their evaluation index of standardized data into the model of mutations, and can obtain their performance scores (table 2). Table 2 shows B through a comprehensive evaluation of the performance of the quality of the highest scores, although the indicators of capacity while the lowest scores, but is no less than with other managers and B in the indicators of performance indicators and evaluation of the attitude of the highest scores. D managers evaluate the performance of the overall quality scores of the second, only slightly higher than B in ability indicators. A manager scores the lowest, although the indicators of the ability and attitude indicators are high, but performance indicators to 0. So B is the best candidate.

TABLE Ⅱ

EVALUATION OF THE OVERALL QUALITY OF CANDIDATE THE PERFORMANCE SCORES

candidate	A	B	C	D	E
$\mathbf{X}_{A_1}$	0.0000	0.8157	0.8061	0.8130	0.7976
$\mathbf{X}_{A_2}$	0.9921	0.9915	0.9933	0.9938	0.9942
$\mathbf{X}_{A_3}$	0.9890	0.9965	0.9987	0.9948	0.9925
$\mathbf{X}_G$	0.6604	0.9346	0.9327	0.9339	0.9281
ranking	5	1	3	2	4

Ⅳ. CONCLUSION

On the use of mutations progression method in this article, enterprise human resources performance has been evaluated and found that the results are reliable. From the calculation process, we can find this method does not need to assign weights of evaluation indexes, and it only consider the relative importance of index, avoid the incertitude and subjective directly use the concept of "large" weight, thus it is simple. This method is applied to the treatment of the performance evaluation of enterprise human resources and should be further refined to promote.

REFERENCES

[1] Zuoping Xiao and Yubao Li, “On the Application of Fuzzy Collection Theory in Performance Appraisal of Human Resource,” Journal of Commercial Research, pp. 1-3, Jul. 2004.

[2] Zhijun Han and Xiaojun Cai, “Application of Analysis of Variance in Human Resources Appraisal,” Journal of Nanjing University of Science and Technology, Vol 27, No.5, pp.541-545, Oct. 2003.

[3] Yan-ling Xiao , Xiaojing Liu , and Jianbo Liu ,“The method of giving weight for performance indicator based

on entropy method," Journal of Daqing Petroleum Institute, vol 29,N0 1,pp. 107 – 109,Feb. 2005 .

[4] Jian Li, " Performance evaluation of the human resource based on ANP," Journal of tianjin university of technology, vol 24,No2,pp.63-66,Apr.2008

[5] Xisong Liu, Chunrong Du, and Yaowu Wang, et al, "Performance appraisal of the human resource management in the construction project," Journal of harbin institute of technology vol 38, No 3, pp. 436-438, Mar.2006.

[6] Xingfu Du, Mutations in the application of theory of the economy, Electronic Science and Technology University Press, Chengdu,1994.

[7] Xiaohong Chen, Jia Peng, and Xiaojin Wu, " Appraisal model based on the sudden change progression for the growth of small and medium- sized enterprises," Journal of Finance and Economics， vol 30,No 11,pp. 5-15, Nov.2004.

[8] Shunquan Zhu, "Study on catastrophe theory and application of credit evaluation of listed corporation," Journal of the System Engineering Theory and Practice, pp. 90-94, Feb.2002.

[9] Yunfeng Chen, Dianyi Sun, and Genfa Lu. "Application of catastrophe progression method in ecological suitability assessment: A case Study on Zhenjiang new area," Journal of Ecological, vol26, No 8, pp. 2587-2593, Aug. 2006.

[10] Kui Wen and Yongsheng Tan, "Try our talents of constructing evaluation index system", Journal of Capital university of economics,pp.5-8,Feb,2005

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 458-461

Image Semantic Classification Using SVM In Image Retrieval

Xiaohong Yu [1], and Hong Liu [2]
[1] College of Computer Science & Information Enginerring , Zhejiang Gongshuang University
No.18,Xuezheng Str.,Xiasha University Town,Hangzhou, China
Email:XHYU@mail.zjgsu.edu.cn
[2] College of Computer Science & Information Enginerring , Zhejiang Gongshuang University
No.18,Xuezheng Str.,Xiasha University Town,Hangzhou, China
Email: LLH @mail.hzic.edu.cn

***Abstract*— There is a gap between low-level descriptions of image content and the semantic understanding of users to query image databases in the content-based image retrieval. In this paper, we put forward a method of classifying image regions hierarchically using their semantics and that resembles peoples' perception more than using low-level features. The experiments show, the better precision of semantic classification justifies the feasibility of our method. It uses in image retrieval field further and get better index effect.**

***Index Terms*— Image classification, semantic classification, image retrieval, Super Vector Machine, keyword-based retrieval**

I. INTRODUCTION

The growth of the World Wide Web have led to the huge online digital images and videos, so there is a strong demand for developing an efficient technique for image retrieval to exploit maximum benefit from this huge amount of digital information. In traditional system, the keywords of image in database are labeled manually and then it utilizes text-based retrieval system to index the image. As the image increases, this technique becomes very inefficient and insufficient to describe the details of an image, so the content-based image retrieval systems have been the major subject for recent decades. Many images retrieval systems have finished, such as QBIC, Visual SEEK, Netra and MARS and so on. They index and retrieval of image based on low-level features of image, such as color, texture and shape. Content-based image retrieval techniques based on similarity matching of features. Take Color similarity for example, color similarity of image can measure by pixel luminance matching, but pixel matching is highly sensitive to noise and small distortions like rotation, further, it is really time-consuming. Most of above systems have the advantage of being automatic, but they are hard to use for novice because of the semantic gap that exists between user perception and system requirements.

As a matter of fact, novice prefer to retrieval image using image semantic elements, such as land, sky, mountain, snow and grass, which are closer to their perception than low-level features. If the system adopt hierarchical semantic to organize and index the image, the gap between the low-level descriptions of image and the user's semantic needs reduce. That is, to reduce the semantic gap, we need to classify image regions based on their semantics. Let novice queries desired images intuitively.

In this paper, we put forward a method of the automatic hierarchical classification of image regions into a more detailed classification hierarchy based on the semantics of the region content by using SVM, and the paper also give a experiment to prove the method that can perform well in the image retrieval field.

We organize this paper as follows. In Sec. 2, we describe the HSV color space that is more suitable for human perception. In Sec. 3, we pay attention to image segmentation, extraction of region features, and simply discuss how to build the SVM and classify the image regions by using SVM. In Sec. 4, the experimental results are presented and finally, the conclusions are given in Sec. 5.

II. HSV COLOR SPACE

Although the process followed the human brain in perceiving and interpreting color is a psychological problem that is not yet fully comprehended. The purpose of color model (also call color space) is to facilitate the specification of colors in some standard. In fact, a color space is a specification of a coordinate system where each color is represented by a single point

Most color space in use today are oriented either toward hardware or toward applications, the hardware-oriented space most commonly used in practice are the RGB (Red, Green, Blue) space, the HSV (Hue , Saturation, Value) space is more suitable for human perceive , so in this paper we use HSV space for studying.

We map the image into the HSV color space.

$$\begin{cases} v = \dfrac{R+G+B}{3} \\ H = \arccos\left\{ \dfrac{\frac{1}{2}[(R-G)+(R-B)]}{[(R-G)^2+(R-B)(G-B)]^{1/2}} \right. \\ s = 1 - \dfrac{3}{R+G+B}[\min(R,G,B)] \end{cases}$$

For each image, we find the minimum and maximum values of each of the three color's components to set up

the coordinate of the HSV space. Each axis runs from minimum to maximum values. These values normalize so that the minimum value equal zero and the maximum value is one. Then, the H, S and V components are quantized to 16, 8 and 8, respectively, within the minimum-maximum range of each component. The H component quantize into more levels, as compared to both S and V, to reflect the diversity of colors in the image database. The values can change by user if necessary to suit a specific image collection.

III. IMAGE CLASSIFICATION BY REGION-BASED ON SEMANTIC

A. Image segmentation

Image segmentation is a process of dividing an image into coherent, uncovered and significative regions. Generic, complete and to the pixel accurate unsupervised segmentation regard as it is virtually impossible, so in this paper, we just want to get a method to segment an image, which is satisfied with the following condition:

(1) First, the extracted regions are coherent.

(2) Segmentation should give satisfactory results on general image data without knowledge assumed.

(3) Segmentation process should be unsupervised.

In our experiment, we use hill-climbing method to segment the image, which can be satisfied with above condition.

The hill-climbing algorithm summarizes as follows:

(1) Compute the HSV color histogram of the image.

(2) Start at a non-zero bin of the color histogram and make uphill moves until reaching a peak.

(3) Choose another unclimbed bin and re-perform step 2 to find another peak. Repeat this step until all non-zero bins of the color histogram climbed.

(4) The peaks we get from above represent the first number of clusters of the input image, and, these peaks saved.

(5) In the end, neighboring pixels that have same peak put together, that is associating every pixel with one of the identified peaks. Consequently, the segments of the input image formed.

The segmentation results shows in Figure 1.

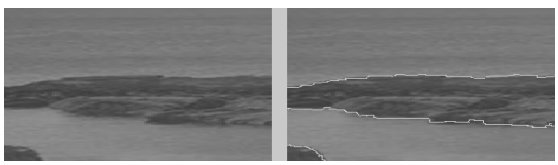


Figure 1. The result of Image segmentation

B. Extract feature of region

We use different way to extract the feature of region.

1) color histogram in the HSV space

In our classification experiments, we found that we could achieve better classification accuracy if we represent the color content of each region with only the H and S components. Thus, we eliminate the V component. A color histogram contains the H component, which is quantized to 6 values, and the S component, which is quantized to 4 values. Hence, the color histogram is represented by a 24-dimensional vector.

2) Edge direction histogram

Edge Direction Histogram of the image shape is one of feature extraction methods, the algorithm extracted feature vectors satisfies scale, translation and rotation change.

There are many different types of edge detector operators. We use the popular Canny edge detector. Experiment proved that the method for a single background, the shape characteristics of clear image with better research results.

C. Support Vector Machine(SVM) Classifier

1) Review of support vector classifiers theory

The way of constructing a hyperplane to get binary classifiers done that can separate members of one class from others, but most real data hardly separate because the hyperplane that can successfully separate the members of the two classes in most case does not exist. One measure to solve this problem is to map the data into a higher dimensional space, where the members of the two classes can separate by a hyperplane. However, the traditional classifier is not good at in high dimensional vector. It is extremely expensive in terms of memory and time.

Support Vector Machines can solve this problem. SVM avoid overfitting the data by choosing a hyperplane from the many that can separate the data. That maximizes the minimum distance from the hyperplane to the closest training point. Such a hyperplane call the maximum margin hyperplane. Another advantage of the SVM is the compact representation of the decision boundary, so the number of support vectors is small as compared to the number of points in the training set.

In this, we simply introduce Support Vector Machine for binary classification

The given training data set for binary classification problem is :

$$\{(x_1, y_1),...(x_i, y_i),...,(x_l, y_l)\} \tag{3.1}$$

where $x_i \in R^d$ and $y_i \in \{-1,1\}$ are training pattern vectors and their corresponding labels, and l indicates the number of training pattern vectors.

Let us also define a linear decision surface by the equation

$$f(x) = w \cdot x + b = 0 \tag{3.2}$$

Where w is normal to the hyperplane, $|b|/\|w\|$ is the distance from the distance from the origin to the hyperplane.

If the following formulation exists:

$$\begin{aligned} w \cdot x + b \geq 1 \qquad & y_i = 1 \\ w \cdot x + b \leq 1 \qquad & y_i = -1 \end{aligned} \tag{3.3}$$

It means the training date set can be separated in linear.

Using a nonlinear transform Φ , these pattern vectors in Eq. (2.1) can be mapped from the original input space R^d into high dimensional feature space R^n , the transform shown in Figure. 2.

$$x \in R^d \rightarrow \Phi(x) \notin R^n \tag{3.4}$$

In the feature space R^n, SVM aims at constructing a linear discriminant function of the form,

$$f(x) = sign(w \cdot \Phi(x) + b) \tag{3.5}$$

Where w and b imply the weight vector and threshold; and <,> denotes the inner product.

According to structural risk minimization principle, SVM is to solve a problem as follows,

$$\min \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \tag{3.6}$$
$$s.t. y_i (w^T \Phi(x_i) + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0 \quad i = 1, \cdots, l$$

Where C is the regularization parameter which can control the tradeoff between the number of errors and the complexity of model, and the slack variable $\xi_i > 1$ corresponds to some misclassified training sample.

$$\max \begin{cases} L_D = \sum_{i=1}^{l} a_i - \frac{1}{2} \sum \sum a_i a_j y_i y_j \Phi(x_i) \cdot \Phi(x_j) \\ = \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} a_i a_j y_i y_j K(x_i, x_j) \end{cases}$$
$$s.t. \quad 0 \leq a_i \leq C \tag{3.7}$$
$$\sum_{i=1}^{l} a_i y_i = 0$$

Where $a_i \geq 0, i = 1, \cdots, l$ are Lagrangian multipliers to solve. $k(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$ is kernel function, Some of the classical SVM kernels are reported in Table 1.

So, the discriminant function and parameter b are:

$$\tilde{f}(x) = sign \sum_{i=1}^{l} y_i a_i K(x_i, x) + b) \tag{3.8}$$

$$b = \frac{1}{N_{NSV}} \sum_{x_i \in JN} (y_i - \sum_{x_j \in J} a_j y_j K(x_j, x_i))$$

Where N_{NSV} is the number of standard support vectors, JN is the set of standard support vectors, J is the set of support vectors.

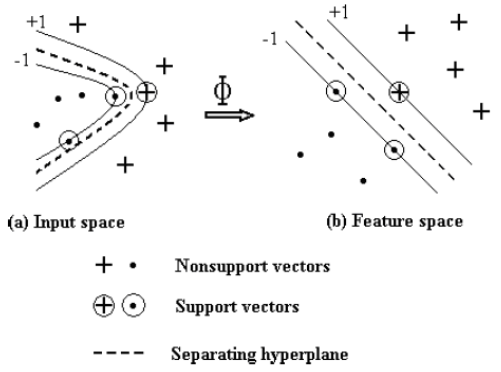


Figure 2. A linear hyperplane separating the members of two classes. The support vectors are circled which is technique used by most radial basis function classifiers.

2) Region Classification

In order to classify image regions into semantic classes in which humans can understand easily, we manually defined a hierarchy organization that reflects the semantics in the Nature images and based on human judgement subjectively, as shown in Figure. 3. The hierarchy organization is not complete by itself, but it is a reasonable organization to simplify image retrieval.

TABLE I.
CLASSICAL COMMON KERNELS (A IN KMOD IS A NORMALIZATION CONSTANT

Kernel	*Formula*
Linear	$k(x, y) = x \cdot y$
Polynomial	$k(x, y) = (ax \cdot y + b)^d$
RBF	$k(x, y) = \exp(-\|x - y\|^2 / \sigma^2)$
KMOD	$k(x, y) = a(\exp(\frac{\gamma^2}{\|x - y\|^2 + \sigma^2}) - 1)$

The selection of classes based on having general concepts to give meaningful associations in normal comprehension.

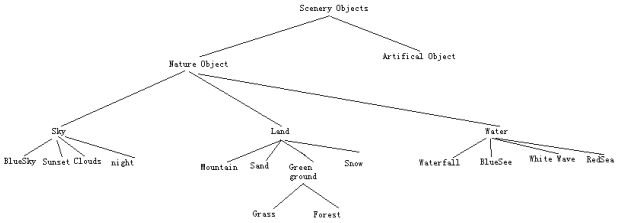


Figure 3. A class organization used the SVM to distinguish the members of a class from others

3) Learning and classification stage

Learning the semantics for each class through using SVM based on different features of training sample regions of each class. We get the SVM classifiers is P(classified | notMemberOfClass) .These binary classifiers achieve good class separation under the constraint that each region belongs to only one, or none, of the classes.

After training the SVM, binary classifiers that can classify image regions based on their semantics create. Then, we use these binary classifiers to classify our database of image regions leading to the classification shown in Figure 3 to determine the class of an input image region and this image region map into its class in the semantic class hierarchy in Figure.3. In the semantic class hierarchy, all scenery regions are classified into Nature regions or Artificial regions. Since in this paper we concentrate on Nature regions, thus we further classify the Nature regions into three subclasses: Sky, Land and Water, Each one of these subclasses further divided into sub-subclasses. The Sky subclass divided into Night, Sunset, Clouds and BlueSky. Next, the Water subclass divided into Waterfall, BlueSea, WhiteWave and River. Then, the Land subclass divided into Mountain, Sand, Greenground, and Snow. The Greenground sub-subclass further divided into Grass and Forest.

Thus, each image can be represented by a set of keywords that are the name of class based on semantic classification of image regions. The choice of keyword-based method allow for highly intuitive query interface,

so the novice can use the semantic to retrieval image by their understanding.

Ⅳ. EXPERIMENTS

We do experiments mainly on images testing set, such as nature scenery, flowers, flags and winter about two thousands images. We divide each image into 5 regions on the average. We got a database with about ten thousands regions as a result of segmentation. We selected 600 regions from above database as a training set for training the SVM. That is about 50 images per class. Then, the extracting feather use color histogram and Edge direction histogram in different classes.

To classify the image regions, we tried different grouping of classes and different features before we finally decided on the grouping and features in Figure. 3, which gives the best classification precision. At is experiment we performed the classification Nature regions and Artifical regions using EDH to extract the feature of region. Then we tried to group the Nature image regions into 3 classes using color histogram feature because it gets high precision than using EDH. Finally, we use EDH feature to group the Water image regions into 4 classes. We get the experimental result shown in Figure. 4 when the user input the keyword waterfall.

Figure 4. Result of the a query of waterfall

Ⅴ. CONCLUSION

We put forward a method to classify image regions based on their semantics. It can reduce the gap between human's perception and description of image content. Because the pre-defined semantic class hierarchy reflects in the semantics by human's subject, so it is flexible and intuitive query by novice.

The use of the binary SVM classifiers that classify image regions using different features at different levels in the hierarchy were the main reasons behind the high classification precision that we achieved in our experiments. Currently, we are looking adding more feature extraction methods to get high precision and put more classifiers to include more classes into the system.

ACKNOWLEDGMENT

This work was supported by Natural Science Foundation of Zhejiang province (No:Y1080565)

REFERENCES

[1] N.E.Ayat, M.Cheiet, C.Y.Suen, Automatic model selection for the optimization of SVM kernels-pattern recognition 1733-1745 (2005)

[2] Zaher Al Aghbari, Region-based semantic image classification –International Journal of Image and Graphics Vol.6. No.3 (3006) 357-375

[3] Pawan Jain, S.N.Merchant, Wavelet-based multiresolution histogram for fast image retrieval. International journal of Wavelets, Multiresolution and Information processing (2004)

[4] M. Flickner, H. Sawhney, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D.Lee, D. Petkovic, D. Steele and P. Yanker, Query by Image and Video Content:The QBIC System, IEEE Computer Magazine (1995).

[5] S. Mehrotra, Y. Rui, M. Ortega and T. S. Huang, Supporting Content-Based Queriesover Images in MARS — Proc. IEEE Int. Conf. On Multimedia Computing andSystems (1997).

[6] Mihalcea, R. and Moldovan, D.: Semantic indexing using WordNet senses. In Proceedings of ACL Workshop on IR & NLP, Hong Kong, October 2000.

[7] Miller, G. Wordnet: A lexical database. Communication of the ACM, 38(11):39--41, (1995).

[8] 7. Joon Ho Lee, Myong Ho Kim, and Yoon Joon Lee. "Information retrieval based on conceptual distance in IS-A hierarchies". Journal of Documentation, 49(2):188{207, June 1993.

[9] Haav, H. M 5. A. Natsev, R. Rastogi and K. Shim, "WARLUS: a similarity retrieval algorithmfor image database," IEEE Transaction on Knowledge and Data Engineering 16(3),(March 2004).

[10] R. Krishnapuram and S. Medasani, "Content-based image retrieval based on a fuzzyapproach," IEEE Transaction on Knowledge and Data Engineering 16(10), (October2004).

[11] F. Jeng, M. Li, H.-J. Zhang and B. Zhang, "An efficient and effective region-basedimage retrieval framework," IEEE Transaction on Image Processing 13(5), (May2004).

[12] J.-H. Lim, "Explicit query formulation with visual keywords," Proc. ACM Multimedia(October 2000).

[13] B. Bradshaw, "Semantic based image retrieval: a probabilistic approach," Proc. ACMMultimedia (October 2000).Region-Based Semantic Image Classification 375

[14] Minka, T.P., Picard, R.W.: Interactive learning using a society of models. Pattern Recogn. 30, 565–581 (1997)

[15] Wood, M.E., Campbell, N.W., Thomas, B.T.: Iterative refinement by relevant feedback in content based digital image retrieval. In: Proceedings of the International Conference on Multimedia, pp.13–20 (1998)

[16] Jing, F., Li, M.J., Zhang, H.J., Zhang, B.: Learning region weighting from relevance feedback in image retrieval. Proc. IEEE Int. Conf. Acoust. Speech Sign. Process. 4, 4088–4091 (2002)

[17] Jing, F., Li, M.J., Zhang, H.J., Zhang, B.: Region-based relevancefeedback in image retrieval. Proc. IEEE Int. Symp. Circ. Syst. 4,145–148 (2002)

[18] Jing, F., Li, M.J., Zhang, H.J., Zhang, B.: Support vector machines for region-based image retrieval. Proc. IEEE Int. Conf. Multimedia Expo. 2, 21–24 (2003)

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 462-466

Suffix Tree Based Chinese Document Feature Extraction and Clustering in RSS Aggregator

Jian Wan, Wenming Yu, and Xianghua Xu
Grid and Services Computing Lab , School of Computer Science and Technology
Hangzhou Dianzi University, Hangzhou 310037, China
[1]yuwenming_001@163.com

***Abstract*—In RSS aggregator, the important issue is how to make the feeds information more manageable for RSS subscriber. In this paper, we propose a suffix tree based RSS feeds document clustering in Chinese RSS aggregator. We construct a suffix tree with meaningful Chinese words, and choose the phrases with high score given by a formula as document features. We cluster document using group-average algorithm with a new document similarity measure. The experiment results show that the new method can improve the quality of clustering in document "snippets" scenario, and the speed can meet the demand of "on the fly" clustering.**

***Index Terms*—suffix tree, feature extraction, document clustering, RSS**

I. INTRODUCTION

RSS (Really Simple Syndication) is a lightweight XML format designed for sharing headlines and other Web content, and provide news updates from a website in a simple form for web user. RSS feeds benefit readers who want to subscribe the timely updates from favorite websites. However, RSS does not solve the problem of information overload for user, and things tend to get worse as the reader subscribes more and more feeds. One way to deal with the information overload problem is to implement a clustering method within a RSS aggregator. By clustering similar items, a feed aggregator can provide a more friendly interface to user, enable the user to quickly filter duplicate or very similar items [1]. It can also help in filtering out topics that the user is not interested in.

Clustering technique relies on four concepts: data representation model, similarity measure, clustering model and clustering algorithm [2]. From all of these parts, the document representation is the most important, because it determines the way that the other three parts choose. In RSS aggregator, Feed usually send a title and a snippet of content for a feed item along with a link to the full content of that item, so the document representation implemented within RSS aggregators needs to extract the document features from the limited RSS snippet content.

In RSS scenario, we choose suffix tree document model [3] which does not treat a document as a bag of words but rather as a string, making use of proximity information between words. By extracting more information present in the documents, we believe suffix tree document model can help in improving the quality of the clusters.

In this paper, we present a novel method for Chinese snippets clustering. Firstly, we obtain meaningful words (always noun and verb in Chinese) from snippets by Chinese word segmentation at the stage of document preparing. In the construction of Chinese suffix tree, we ignore the nodes (feature phrases) with a high document frequency (*df*), and only choose the nodes with high score given by a formula we proposed. Then we redefine the pair-wise documents similarity measure for RSS snippet content. With combination of the document features extracted based on suffix tree and the new document similarity, the group-average AHC algorithm is realized in RSS aggregator. The experiment results show that the new method can improve the quality of clustering, and the speed can meet the demand of "on the fly" clustering.

The rest of the paper is presented as follows. Section 2 discusses related work. Section 3 present the improved clustering approach which can be used in Chinese RSS aggregator. Section 4 illustrates some experimental results. Finally Section 5 summarizes our work with some considerations on future directions.

II. RELATED WORK

Document clustering has been investigated as a post-retrieval document browsing technique. Most clustering algorithms base on two document model: the vector space document (VSD) model[4] and the suffix tree document model[3]. Clustering method base on VSD model such as K-Means and agglomerative hierarchical clustering (AHC) cluster the document according to the similarity of vectors which represent documents in the defined vector space. There are several variants from AHC[5], e.g. single-link, group-average, and complete-link. These original algorithms are usually too slow to meet the requirement of "on-line" web applications, such as RSS feeds stream clustering application.

Suffix tree clustering (STC) Algorithm based on Suffix tree document model are usually used in English search's results clustering[3, 6]. STC is an incremental clustering algorithm and its time complexity is linear with regard to the document corpus size, so it is suitable for clustering web document snippets returned from search engine. However the clustering effectiveness of STC is unsatisfied in Chinese RSS Snippets[7].

Recently, many clustering methods are extended to

AP-PROC-CS-09CN005

specific domains to make the information more manageable. Chim and Deng[8] have proposed a new clustering algorithm combine the advantages of two document models in document clustering. Peng Jing[9] present a novel Chinese text clustering algorithm based on inner product semantic space model. There methods can improve the clustering quality, but they are suitable for "off-line" clustering situation due to time efficiency. Some special methods such as extracting feature code and compressing code[10] are also proposed to solve the clustering problem of short documents.

Compared to aforementioned work, the new clustering algorithm we proposed is to improve the quality of STC in clustering Chinese snippets, and the clustering speed can meet the demand of "on the fly" mode in RSS Aggregator.

III. A New Clustering Algorithm in Chinese Snippets Context

Our method has three logical steps: (1) document preparing, (2) extracting key phrase using a suffix tree, and (3) clustering snippets using group-average algorithm with a new document similarity measure.

A. Document Preparing

We would ideally like to do the clustering "on the fly" within the RSS aggregator, this means that we do not have time to download the complete content. So we take a title and a snippet of a feed item as a good summary of its content, and treat them as a "document" (document we used in the following means the title and the related snippet) to be clustering. There documents are parsed and split into sentences according to punctuations and HTML tags, and all empty words are stripped. However, different from English document, Chinese words are base units in Chinese document from the view of semantics. Therefore we incorporate Chinese word segmentation into Document Preparing.

Chinese words are fewer than Chinese characters in the same document, and the nodes in suffix tree based on Chinese words can be fewer than on Chinese characters. This means speeding up the construction of suffix tree. Furthermore, the meaningless nodes are removed to improve the accuracy of clustering results. For example, if the Chinese phrase "中山广场" (ZhongShan Square) is identified as a phrase, then its sub-string "山广场" (Shan Square) is inevitably selected as a phrase, but this phrase is meaningless [7].

Part-of-Speech (PoS) selection always combines with Chinese word segmentation. In Chinese language, the empty words such as adverb, adjective, preposition, and conjunction act as modifier, and have little power of discrimination. On the other hand, the same semantic snippet can use different empty words based on different context. We can't extract document features effectively if don't remove the empty words which affect identifying of the common phrase. So we only reserve noun and verb for clustering, this can also reduce nodes in a suffix tree.

B. Feature Extraction

We decide to use key phrases extracted from the document collection based on suffix tree as document features. We believe this can help in improving the quality of the clusters by leveraging more information present in the documents.

a. Construct Suffix Tree

The suffix tree data structure was introduced as an efficient string processing technique. A suffix tree allows us to insert a string into the suffix tree incrementally. Following is the definitions related to a suffix tree which was built with Chinese words.

Definition (Suffix of Chinese String): Suppose a string (sentence in document) $d = w_1 w_2 \cdots w_m$ consists of Chinese noun or verb $w_i (i = 1, 2 \cdots m)$, then $S_i = w_i w_{i+1} \cdots w_m$ is a suffix of d starts from the position of i.

Definition (Suffix Tree Chinese Document Model): The suffix tree of Chinese documents is a compact tree containing all the suffixes of sentences in documents (designated by leaf nodes). Each edge is labeled with a non-empty substring of sentences. No two edges out of the same node can have edge-labels that begin with the same Chinese word. Each internal node has at least 2 children, represents an overlap phrase shared by at least two suffixes.

Definition (phrase): Phrase is the label of a node, which is designated by the concatenation of the edge-labels on the path from the root to that node.

Figure 1 is an example of a suffix tree composed from 3 simple documents which have processed by document preparing. The 3 processed documents are '猫 吃 老鼠' (Cat ate mouse), '老鼠 吃 肉' (Mouse ate meat) and '猫 吃 肉' (Cat ate meat). The nodes of the suffix tree are drawn in circles. Each internal node is attached with a box respectively, each upper number designates a document identifier that presents which document have traversed the corresponding node, each below number designates the phrase occurs in title or in body (0 in body, 1 in title, no title in our 3 example documents).

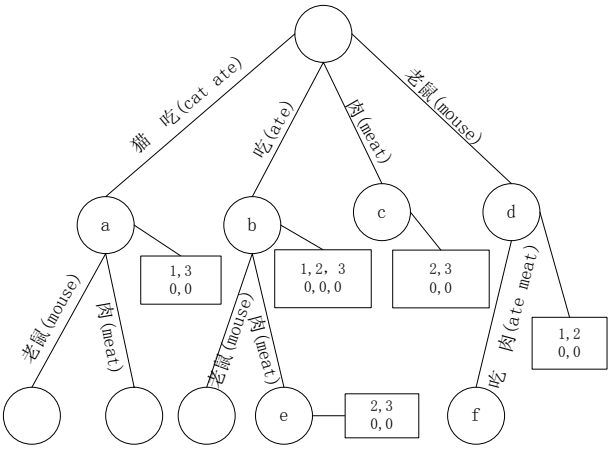


Figure 1 The suffix tree of three documents

b. Feature Extraction

The suffix tree constructed from documents usually contains lots of (about ten thousand) internal nodes (phrases). Not all internal nodes (phrases) are useful for document clustering, and some of the nodes (phrases) may even misguide the clustering results. So selecting a subset of original nodes as document features can not only reduce the high dimensionality of the feature space, but also improve the accuracy of clustering results.

Although the original suffix tree clustering evaluate the quality of nodes with an empirical formula. However, it tends to choose internal nodes containing more documents in our experiment. This means the phrases designated by these nodes have little power of discrimination.

We removed nodes with too high document frequency (df) using a threshold ($df - threshold$). For example, Many news contains the phrase "记者从相关部门获悉" (The reporter learned from relevant departments that). Although the sub-phrase "记者" (reporter) can express specific meaning in different context, which means we should reserve it. However, we ignore the whole phrase with high df in the document similarity measure.

Then, we evaluate the phrase importance by statistical method which is usually applied in VSD model. For each internal node n_i, the phrase designated by n_i is p_i. When we are constructing the suffix tree, use variable $tf(p_i)$ accumulate the times traverse through the node n_i by the suffix in the documents corpus, then $tf(p_i)$ is the term frequency of phrase p_i; the times of different documents that traverses through the node n_i is $df(p_i)$, then $df(p_i)$ is the document frequency. Therefore the weight of phrase p_i in documents corpus can be calculated using the classic tf/idf scheme in formula (1), where N is the total number of documents in corpus.

$$tf/idf = \log(tf(p_i)) \cdot \log(N/df(p_i)) \qquad (1)$$

A phrase is an ordered sequence of one or more words. The more number of words a phrase contains, the richer meaning it can express. Therefore, the importance of a phrase should incorporate a factor about the length of phrase p_i, which designated by $|p_i|$. We calculate the factor using a heuristic utility function in following formula:

$$f(|p_i|) = \log_2 |p_i| \qquad (2)$$

The score $s(n_i)$ of node n_i (phrase p_i) is given by formula (3). To speed up the follow clustering, we only choose the k highest scoring phrases as key phrases (we take k to be 1000 in our experiment).

$$s(n_i) = tf/idf \cdot f(|p_i|) \qquad (3)$$

C. Document Clustering

The document clustering approach we presented is an improved group-average algorithm[5] in which the pair-wise documents similarity measure is modified according to RSS context. The advantage of group-average clustering algorithm in our application is it can be stopped at any point when the remaining pairs of clusters have low similarity values, so it doesn't have the problem of selecting appropriate initial number of clusters. Most importantly, group-average algorithm can always achieves better clustering result than other algorithms.

Group-average algorithm is argued for spending too much time in clustering. But in our clustering problem, the number of clusters k tends to be comparable to the number of documents n. This is because there are few similar news about the same topic, majority of clusters have only several news each. Therefore, the group-average algorithm with time efficiency $O(n^2)$ can be faster than EM or K-Means implementations, which is $O(knf)$ (where f is the number of features per document) per iteration.

We redefine the pair-wise documents similarity abandoning the cosine document similarity measure to speed up the Clustering. From the suffix tree constructing, it's very easy to understand that the more internal nodes shared by two documents, the more similar the documents tend to be. Since the key phrase in the title of a document is more representative than in the body, it is reasonable to distinguish different situation with different weight. So we use the doc_similarity algorithm (shown in Figure 2) to measure the similarity between two documents.

```
1:  void doc_similarity(sorted array doc1,
                                sorted array doc2)
2:  {
3:      int  i = 0;
4:      int  j = 0;
5:      int  similarity = 0;
6:      while ((i<k) && (j<k)){
7:          if( doc1[i] == doc2[j] ){
8:              if( doc1[i].pos==1 &&
                                doc2[j].pos==1)
9:                  similarity += 5;
10:             else if(doc1[i].pos==0 &&
                                doc2[j].pos==0)
11:                 similarity += 3;
12:             else
13:                 similarity += 1;
14:             i++;
15:             j++;
16:         }
17:         if(doc1[i] < doc2[j])
18:             i++;
19:         if(doc1[i] > doc2[j])
20:             j++;
21:     }
22: }
```

Figure 2 The function to measure the similarity between two documents

Each document can be expressed with an array related to the k key phrases if we traverse the suffix tree once. We sort each array to make it ordered before measuring the similarity.

IV. EVALUATION

In this section, we evaluate the effectiveness and efficiency of our algorithm. The algorithms to be compared are the original STC and group-average algorithm with the traditional term tf/idf cosine similarity measure. We use JAVA as the tool for simulation, and our experiment equipment is a PC with Pentium(R) 4 CPU 3.00GHz, 1024 MB memory and MS Windows XP operating system.

A. Dataset

Google News service(http://news.google.cn/) uses document clustering techniques to group news articles from multiple news sources, it provide an easily available labeled dataset for evaluating our clustering results.

Each article in Google News homepage has a link pointing to the other articles on the same topic. We choose 8 articles from different topics (e.g. business, sports, health, Entertainment, etc.) and collect the top 100 related articles for each topic. This way we gathered 800 new articles belonging to the 8 different topics. Each news article consists of a title and a description snippet of content.

B. Quality measure

We use commonly used *F-measure* for evaluating and comparing different clustering results. *F-measure* combines the Precision and Recall ideas from the Information Retrieval literature. The precision and recall of a cluster j with respect to a "correct" class i are defined as:

$$P = Precision(i, j) = \frac{N_{ij}}{N_j} \quad (4)$$

$$R = Recall(i, j) = \frac{N_{ij}}{N_i} \quad (5)$$

where

N_{ij} : is the number of members of "correct" class i in cluster j,

N_j : is the number of members of cluster j, and

N_i : is the number of members of "correct" class i.

The *F-measure* of a class i is defined as:

$$F(i) = \frac{2PR}{P+R} \quad (6)$$

C. Results

Although group-average can be easy to tune the similarity threshold to get appropriate clusters in real RSS aggregator, we select 8 initial clusters beforehand to achieve a fair result.

Figure illustrates the *F-measure* scores computed from three clustering algorithms on the dataset we get above. NSTC designates the algorithm we describe above, and GTC designates group-average algorithm with traditional term tf/idf cosine similarity measure.

As shown in Figure 3, the performance of NSTC is improved compared with the two other algorithms. This is mainly due to three facts: (1) We construct the suffix tree based on Chinese words rather than Chinese characters. (2) We extract high discrimination phrases from documents, and remove the meaningless phrases which misguide the clustering result. (3) Our new document similarity measure has the ability to accurately judging the relation between snippets.

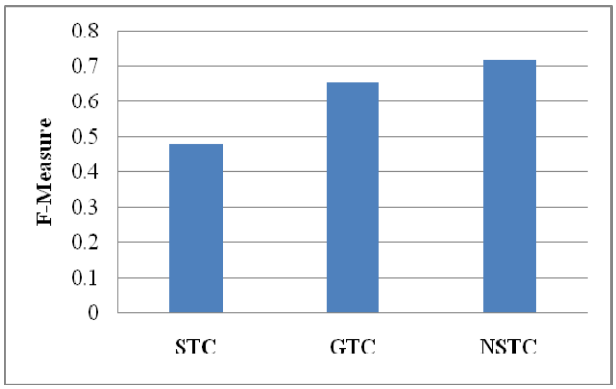


Figure 3 *F-measure* of the three clustering algorithms results

Speed also plays an important role in the implement of clustering for RSS aggregator. We measure the execution time of the different algorithms while clustering snippet collections of various size (100~800 snippets, evenly distributed in 8 different topic as far as possible). The results are shown in Figure 4.

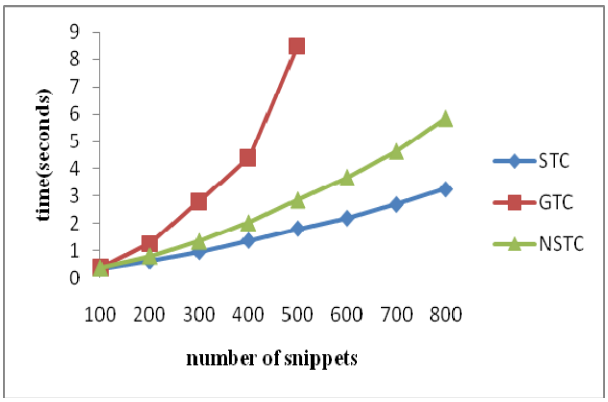


Figure 4 Execution time (in seconds) of different algorithms while clustering snippet collections of various sizes

Although NSTC is much faster than the GTC, it is slower than STC algorithm. The main reason is that group-average algorithm wastes a little time. However, the number of snippets to clustering in RSS aggregator is no more than 400 in most cases. So the efficiency can meet the demand of on-line interaction in RSS context.

V. CONCLUSION AND FUTURE WORK

In this paper, a new document clustering algorithm is introduced for the Chinese snippets in RSS aggregator. We extract effective phrases from snippets based on

suffix tree, and clustering RSS snippets using group-average algorithm with a new document similarity measure. The experiment results show that the new method can improve the quality of clustering, and the speed can meet the demand of "on the fly" clustering in RSS aggregator. In future work, we intend to introduce world knowledge such as HowNet (An electronic dictionary of Chinese like WordNet) to improve the effectiveness of our short Chinese document clustering method, and applying this snippets clustering into other domains such as email clustering.

ACKNOWLEDGEMENTS

This paper is supported by National Science Foundation of China under grant No.60873023, and Science and Technology R&D Program of Zhejiang Province, China under grant No. 2008C13080, No.2007C21G3230005.

We thank Dawid Weiss and Stanislaw Osinski for their contributions to open source framework of clustering (http://project.carrot2.org/), by referring which we reduce amount of time for our research. Thanks also to Institute of Computing Technology, Chinese Academy of Sciences for their excellent work in the implementation of Chinese word segmentation (http://ictclas.org/). We make use of ICTCLAS module in the stage of document preparing.

REFERENCES

[1] Somnath, B., R. Krishnan, and G. Ajay, Clustering short texts using wikipedia, in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. 2007, ACM: Amsterdam, The Netherlands.

[2] Hammouda, K.M. and M.S. Kamel, Efficient phrase-based document indexing for Web document clustering. IEEE Transactions on Knowledge and Data Engineering, 2004. 16(10): p. 1279-1296.

[3] Oren, Z. and E. Oren, Web document clustering: a feasibility demonstration, in Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. 1998, ACM: Melbourne, Australia.

[4] Salton, G., A. Wong, and C.S. Yang, A vector space model for automatic indexing. Commun. ACM, 1975. 18(11): p. 613-620.

[5] Peter, W., Recent trends in hierarchic document clustering: a critical review. Inf. Process. Manage., 1988. 24(5): p. 577-597.

[6] Oren, Z. and E. Oren, Grouper: a dynamic clustering interface to Web search results, in Proceedings of the eighth international conference on World Wide Web. 1999, Elsevier North-Holland, Inc.: Toronto, Canada.

[7] Jiangning, W. and W. Zhijiang. Search Results Clustering in Chinese Context Based on a New Suffix Tree. in Proceedings of IEEE 8th International Conference on Computer and Information Technology Workshops, 2008.

[8] Hung, C. and D. Xiaotie, A new suffix tree similarity measure for document clustering, in Proceedings of the 16th international conference on World Wide Web. 2007, ACM: Banff, Alberta, Canada.

[9] PENG Jing, et al., A Novel Text Clustering Algorithm Based on Inner Product Space Model of Semantic. CHINESE JOURNAL OF COMPUTERS, 2007. 30(8):p. 1354—1363

[10] HUANG Yong-guang, et al., A Fast Clustering Algorithm for Abnormal and Short Texts. JOURNAL OF CHINESE INFORMATION PROCESSING, 2007(02).

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 467-470

Comprehensive Evaluation of Working Environment under Mining Based on Unascertained Analytic Hierarchical Model

Xuanchi Zhou[1], Fengping An[2], and Jun'e Liu[3]
[1] Postgraduate Department of Beijing WUZI Universityt, Beijing,China
Email: zoezxc@126.com
[2] School of Economics and Management of Hebei University of Engineering,,handan,China
[3]Information School of Beijing WUZI University,Beijing, China
Email: anfengping1985@163.comr, zl-je@163.com

***Abstract*—Working environment under mining plays an important role in safety of coal mines. The paper analyzes the eight factors of working environment under mining: gas density, dust, temperature, humidity level, air velocity, and harmful air, noise, and working space, calculates the weight of factors and establishes the comprehensive evaluation model with analytic hierarchical model based on unascertained measure. This paper develops the module by the use of VC++6.0 and SQL Server2000 which can display the gas density curve and Evaluation curve with the help of ProEssentials. This method has been proved usable and reasonable by using the example. And there will be a promising future in its application.**

***Index Terms*—Unascertained measure; AHM; working environment under mining; ProEssentials**

In human-machine-environment system, environment is a significant factor which can deeply affect the safety of a system. Working environment underground directly affects the efficiency of human-machine system, mental and physical health and safety of operators as well. Those situation that large numbers of harmful air released in the course of coal production, rising of air temperature along with the increase of temperature of surrounding rock, high humidity level, large amounts of dust and noise source, and narrowness of working space, will easily cause fatigue for workers, which are key threats to safe production of coal mining.

Synchronously, gas density also plays an important role in the safety of working environment under mining. With the increase of gas density, oxygen density will decrease, which cause workers suffer oxygen-poor and suffocation. Explosion will take place when gas density is beyond the limitation and confronted with origin of heat which is high temperature, and all these will bring about injuries and deaths. Different coal mining face has different limitation of gas density according to coal mine safety rules[1].Take coal mining face for an example, computing formula of reliability of gas density is Y=1-x/0.2.The real-time value of x can be gotten from sensor monitoring system, then it can compute the value of index. In the actual project, in order to detect and assess the influence of working environment under mining on coal mine safety, this article achieves the comprehensive evaluation of working environment under mining which also has been proved more reasonable and directviewing by utilizing Unascertained and AHM theory and using ProEssentials.

I. INTRODUCTION TO PROESSENTIALS

ProEssentials is the product of Gigasoft. Gigasoft is a company develops charting components in USA and provides custom programmed charting solutions to the world's leading companies, including IBM, Microsoft and so on[4].

GigaSoft ProEssentials is a set of charting components for Windows client-side and server-side development. It comes with NET, DLL, ActiveX, VCL, WinForms, and WebForms interfaces which provide convenience for developer to apply in a variety of development with Visual Studio. NET, VC6, VB6, ASP, ASP.NET, Delphi, Builder.

ProEssentials consists of five charting components: Graph，Scientific Graph、3D Scientific Graph，Polar，Pie Chart and that realize the 2-dimention & 3-dimention graphics functions using Cartesian's coordinate system, Polaris/Smith/Rose Char and Pie Char under the polar coordinates system.

2-dimention graphics function can be realized under linearity and logarithmic coordinates system. Methods of drawing include point, line, bar, area and contour, and it can also create shadow and 3D effects, display a variety of bitmap types: JPEG, PNG and BMP, support print output, message and events mechanism which make it convenient for users to interact with the displayed data directly

There are three kinds of data displayed in graphics type: (1) Y= {y1, y2, ..., yn} ; (2) Y= f(x); (3) Y=f(x, z) in normal application.(1) is array or set,(2) and (3) are continuous functions. ProEssentials use Graph ， Scientific Graph and 3D Scientific Graph to express these three types of graphic images. The term of variable Y in above formula is subset. One graph can have six subgraph at the most. Every subgraph includes two ordinate axis (y axis and right y axis) and two cross shafts(x axis and top x axis).The scale of axis can be adjusted on the basis of inputted data automatically or be set up in the program in advance.

II. EVALUATION INDEX SYSTEM OF WORKING ENVIRONMENT UNDER COAL MINES

A. The Establishment of Indexes

Working environment under coal mines is a set of correlated influencing factors concerning comfortable quality, working efficiency and system reliability within the space of coal mining face.

Based on the definition and characteristics of working environment under coal mines, and on the principle of scientific nature, systematic nature, comparability and operability, we conclude multi-level comprehensive index of working environment under coal mines: gas density, dust, temperature, humidity level, air velocity, and harmful air, noise, and working space.

B. The Division of Evaluation Grade

In this article, working environment under coal mines is divided into four grades: very safe, safe, dangerous, very dangerous.

III. UNASCERTAINED ATTRIBUTE AHM

set $x_1,x_2, \cdots ,x_n$ as n objects for evaluation, then $X=\{x_1,x_2, \cdots, x_n\}$as evaluation object space; Each object of study has m kinds of attribute $I_1,I_2, \cdots,I_m$ which can be measured; $I= \{I_1,I_2, \cdots,I_m\}$ are attribute space. x_{ij} is evaluation value of x_i on I_j. Evaluation value x_{ij} could be calculated, so evaluation matrix $(x_{ij})_{n\times m}$ is known. Line i in this matrix expresses observed value of object i on m kinds of attribute, i=1,2….,n; Row j expresses observed value of various objects on attribute I_j, j=1,2….,m.

For every x_{ij}, we can calculate the μ_{ijk} which represent the grade of object x_i on c_k (k=1,2,⋯,K); the process above is also calculating the grade Evaluation of x_{ij} on every c_k. set c_k represents the grade of project risk, the grade K is prior to the grade K+1.If $\{c_1,c_2,\cdots, c_k\}$ accords with $c_1>c_2>\cdots>c_k$, $\{c_1,c_2,\cdots, c_k\}$ is called a ordered division genus of evaluation space U.

A. Single Index Recognition

For every single factor index (attribute) I_j (j=1,2,⋯,m), x_{ij} is given (i is solid). Calculating the measure of x_i that has observed value x_{ij} on c_k (k=1,2,⋯,K) grades is equal to calculating the grade measure of observed value x_{ij} on c_k.

Conforming the measure function $\mu_{ij}(x)$ and calculating the μ_{ijk} for every quality grade k (k=1,2,⋯,K), we can get the Unascertained Measure recognition matrix under the single index:

$$\mu_i = \begin{pmatrix} \mu_{i11} & \mu_{i12} & \cdots & \mu_{i1k} \\ \mu_{i21} & \mu_{i22} & \cdots & \mu_{i2k} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{im1} & \mu_{im2} & \cdots & \mu_{imk} \end{pmatrix} = \left(\mu_{ijk}\right)_{m\times n} \qquad (1)$$

Thereinto, the line t expresses the measure of object xi belongs to each quality grade about the tth kind of observed value; The rows expressed measure that xi belongs to the sth quality grade about various attribute observed value.

B. The Essentiality Weight of Every Index Determined by Applying AHM

(1) Introduction of AHM

AHM is a non-structure decision method, and it is improved from AHP. Compared with AHP, AHM is much easier to do. AHM does not need to calculate eigenvector, and it does not need to check the consistency. It only needs to make multiplication and addition operation.

(2)Steps of calculating weight based on AHM

1) There are many influence factors of working environment under coal mines, so it needs many experts to participate in the evaluation. The basic idea is: firstly evaluate the index's importance on each level separately by the experts; finally the experts calculate the arithmetic average of the index in corresponding level to get the synthetic evaluation result. In identical level, the various indexes get corresponding importance by comparing.

Suppose that there are n factors $b_1,b_2,\cdots, b_n$, if the importance of b_i are the same as the importance of b_j, then $b_{ij}=1$;if b_i is slightly important than b_j, then $b_{ij}=3$; if b_i is obviously important than b_j, then $b_{ij}=5$; if b_i is more important than b_j, then $b_{ij}=7$;if b_i is absolutely important than b_j, then $b_{ij}=9$.Between them there are $b_{ij}=2,4,6$ or 8.It is obvious that $b_{ij}=1/b_{ji}$.

2) Transforms 1-9 scale judgment matrix into AHM, and the transformation procedure is as follows:

$$\mu_{ij} = \begin{cases} \dfrac{2k}{2k+1} & a_{ij} = k \\ \dfrac{1}{2k+1} & a_{ij} = \dfrac{1}{k} \\ 0.5 & a_{ij} = 1 \quad i \neq j \\ 0 & a_{ij} = 1 \quad i = j \end{cases} \qquad (2)$$

It is Obvious that $\mu_{ij}=0$，$\mu_{ij}\geq 0$，$\mu_{ji}\geq 0$，$\mu_{ij}+\mu_{ji}=1(i\neq j)$, μ_{ij} is called the measure under AHM. When $\mu_{ij}\geq \mu_{ji}$, it means that the plan P_i is better than the plan P_j.

3) Make that

$$f_i = \mu_{i1}+\mu_{i2}+\cdots+\mu_{in} = \sum_{j=1}^{n}\mu_{ij}(i=1,2,\cdots;n) \qquad (3)$$

$$c_i = \frac{2f_i}{n(n-1)} \qquad (4)$$

c_i expresses the score rate of μ_i, then $c=(c_1,c_2,\cdots, c_n)$ and $\sum_{i=1}^{n} c_i = 1$. According to the above, the place of each plan can be calculated, named essentiality order.

C. Identified Rule

A confidence threshold is pre-determined called λ (λ>0.5). According to the background and needs of the problem, λ is normally be admitted between 0.6 and 0.8, if $F_i>F_{i+1}$,$\{F_1, F_2,\cdots, F_k\}$ is ordered division ,then

$$k_0 = \min_k\left(k : \sum_{l=1}^{k}\mu_{il} \geq \lambda \ , \ 1\leq k\leq K\right) \qquad (5)$$

Sample x_i belongs to k_0 genus F_{k0}, and the confidence is λ.

The implication is that: the confidence that grade of x_i is not higher than F_{k0} is λ or the confidence of that grade of sample x_i is higher than k_0+1 is 1- λ.

IV. EXAMPLE ANALYSIS

Table 1 is evaluation indexes of working environment under coal mines according to the site data of coal mine in Shanxi province. There are eight evaluation indexes in table 1.

TABLE I. EVALUATION INDEXES OF WORKING ENVIRONMENT UNDER MINING

working environment under coal mines I							
gas density	dust	temperature	humidity level	air velocity	harmful air	noise	working space

A. *Determination of Evaluation Index System of working environment under coal mines*

The evaluating index system of working environment under coal mines is shown in Table 1.

B. *Determination of Various Factors' Weight based on AHM*

Synthesizing opinions of the fellow experts and technician, obtains the importance comparison matrix during various factors concerning working environment under coal mines is obtained:

$$R = \begin{bmatrix} 1 & 8 & 7 & 6 & 9 & 6 & 9 & 9 \\ 1/8 & 1 & 1/3 & 1/2 & 2 & 1/4 & 3 & 3 \\ 1/7 & 3 & 1 & 2 & 4 & 1/2 & 4 & 4 \\ 1/6 & 2 & 1/2 & 1 & 2 & 1/3 & 3 & 3 \\ 1/9 & 1/2 & 1/4 & 1/2 & 1 & 1/5 & 2 & 2 \\ 1/6 & 4 & 2 & 3 & 5 & 1 & 5 & 5 \\ 1/9 & 1/3 & 1/4 & 1/3 & 1/2 & 1/5 & 1 & 1 \\ 1/9 & 1/3 & 1/4 & 1/3 & 1/2 & 1/5 & 1 & 1 \end{bmatrix}$$

Transform it to the judgment matrix under AHM by using (2),

$$R' = \begin{bmatrix} 0 & 0.941 & 0.933 & 0.923 & 0.947 & 0.923 & 0.943 & 0.943 \\ 0.0588 & 0 & 0.143 & 0.200 & 0.800 & 0.111 & 0.143 & 0.143 \\ 0.0667 & 0.857 & 0 & 0.800 & 0.889 & 0.200 & 0.889 & 0.889 \\ 0.0769 & 0.800 & 0.200 & 0 & 0.800 & 0.143 & 0.857 & 0.857 \\ 0.0526 & 0.200 & 0.111 & 0.200 & 0 & 0.0909 & 0.800 & 0.800 \\ 0.0769 & 0.889 & 0.800 & 0.857 & 0.938 & 0 & 0.938 & 0.938 \\ 0.0526 & 0.143 & 0.111 & 0.143 & 0.200 & 0.0625 & 0 & 0.500 \\ 0.0526 & 0.143 & 0.111 & 0.143 & 0.200 & 0.0625 & 0.500 & 0 \end{bmatrix}$$

By Using (3) and (4), we can get:

W'=(4.64,1.61,6.21,9.07,2.26,5.45,0.72,0.72)

The unitary weight is:

W=(0.23,0.06,0.17,0.14,0.11,0.19,0.05,0.05)

C. *Inviting Many Experts to carry on the Risk Evaluation, the Various Indexes Evaluation Matrix are Obtained, they are as follow:*

$$I = \begin{bmatrix} 0.4 & 0.3 & 0.2 & 0.1 \\ 0.5 & 0.3 & 0.1 & 0.1 \\ 0.4 & 0.2 & 0.2 & 0.2 \\ 0.5 & 0.2 & 0.1 & 0.2 \\ 0.6 & 0.2 & 0.1 & 0.1 \\ 0.4 & 0.2 & 0.2 & 0.2 \\ 0.3 & 0.3 & 0.2 & 0.2 \\ 0.5 & 0.2 & 0.2 & 0.1 \end{bmatrix}$$

So,

$W_{comprehensive}$=(0.442,0.234,0.169,0.155)

D. *Recognition, taxis and actual measurement and analysis with ProEssentials*

In this article the grade of working environment under coal mines is divided into four grades: "very safe, safe, dangerous, very dangerous", which is ordered division, therefore we use the confidence criterion, and make the confidence λ=0.6, and the identified matrix of single index recognition measure can result in the evaluation result: the working environment under coal mine is safe.

Actually, ProEssentials provide five interfaces for developer to use. The paper choose VC++6.0 [6] [7] and SQL Server 2000[8] as tools, with database technology to develop application software.

Figure 1 is Evaluation curve by use of ProEssentials. Lateral axis is using Scientific Graph to display time axis, longitudinal axis is realized with Scientific Graph. Left-longitudinal axis expresses gas density; right-longitudinal axis indicates comprehensive evaluation index number of working environment under mining. The scale of axis is set up in the program.

As figure 1 shown, active line states gas density, short dash line expresses index number of risk evaluation. Gas density is the most significant factor result that the program display gas density specially. It can be seen in figure 1 that comprehensive evaluation index number of working environment under mining is 87.5 which indicate that working environment under coal mines at present is in safe state.

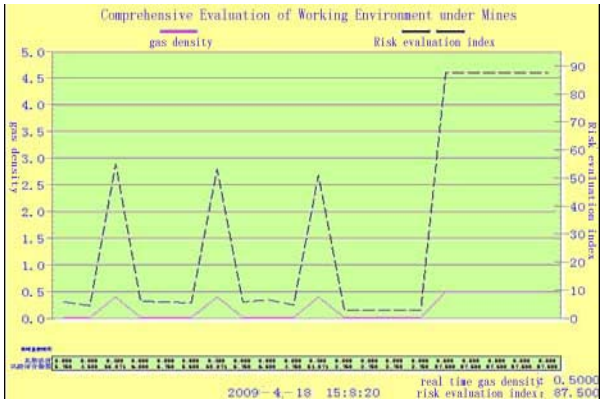


Figure 1. Evaluation curve.

Ⅴ. CONCLUSION

Integrating unascertained and AHM theory, this article calculates the weight of factors with analytic hierarchical model based on unascertained measure and then applies the unascertained theory to working environment under coal mines, achieves the comprehensive evaluation finally that is creative to some extent. The unascertained measure method pays attention to" the ordered nature" of the evaluation space , and makes the reasonable confidence identify and taxis grade criterion, which make the evaluation result more clarity and more reasonable. In this article, the structure decision theory and the non-structure decision theory have been effectively unified. Synchronously, it realizes the combination of a qualitative analysis and quantitative analysis. Finally, this method has been proved applicable and reasonable with the example. This paper develops the module by VC++6.0 and SQL Server2000 which can display the gas density curve and Evaluation curve with ProEssentials. It will be widely used in the future.

ACKNOWLEDGMENT

It is a project supported by Disciplines and Postgraduate Education of Beijing WUZI University, and Beijing Municipal Project for Developing Advanced Human Resources for Higher Education.

REFERENCES

[1] State Administration of Work Safety, State Administration of Mine Safety. Safety regulations in coal mine [M]. Coal Industry Press,2006 .

[2] CHENG Qiansheng. Analytic Hierarchy Process (AHP) and Attribute Hierarchical Model (AHM)[J].System engineering-Theory &practice, 1997, (11).56-59.

[3] XUE Hua. Analysis on Disaster of Coal Mine[J] .Coal Technology.2009.2.95-96

[4] Gigasoft Company. ProEssentials home page[EB/OL].http://www.gigasoft.com.

[5] CAO Qingkui, Wei Xiaoling, Hu Jinsheng,Liu Zhiqiang. Decision model of machine design scheme [J].Journal of Mechanical Strength 2003,25(4):470-472.

[6] WANG Yu-han. Four Methods to Realize Dynamic Curve by Visual C++ Programming [J].Journal of Chongqing Institute of Technology. 2008.6 (6):94-98.

[7] (USA) Kruglinski DJ. visual c++ 6.0 technology insider[M].Beijing: Tsinghua University Press,1999.

[8] SONG Kun, LI Yan, WANG Guohui. SQL Server database development classic case analysis [M].Beijing: Machinery Industry Press.2006

[9] Saaty T L. The Analytic Hierarchy Process [M]. Pittsburgh: university of Pittsburgh, 1988.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 471-474

Organizing and Implementing Method on Joint Combat Experiment

Yanfei Han
Dalian Naval Academy/Missile Department, Dalian, China
Email: sonyhyf@sina.com

***Abstract*—The paper studied meta-synthesis theory and applied it to the field of combat experiment. Firstly, the paper divided experiment modes and experiment phrases. Then it divided experiment roles, experiment duties, and analyzed organization of small combat experiment. At last, it pointed out organizing and implementing methods of large joint combat experiment. All the ideas in this paper had been realized in computer network. The paper had top-floor traction functions and had been used to guide combat experiment development.**

***Index Terms*—joint combat experiment, organization, implementation, method**

I. Introduction

Methods of organization and implementation of joint combat experiment are important study results of meta-synthesis theory applied into the field of application of joint combat experiment. Part of study results are selected and form this paper. In this paper, joint combat experiment modes, phrases, roles, means and methods of organization and implementation are analyzed. Finally, an example of joint combat experiment is given.

I I. Joint Combat Experiment Modes and Phrases[1]

A. Joint Combat Experiment Modes

Individual Study Mode: Individual study is a basic mode of joint combat experiment. It refers to single consumer or single person's experiment activities on a platform of joint combat lab. In this mode, each experimenter goes about his experiment such as information consultation, information analysis, military mark, combat calculation, combat plan editing, combat simulation and evaluation, etc.

Collective Discussion Mode: Collective discussion is a main mode of joint combat experiment. This mode is based on individual study mode. In this mode, experimenters may go about their experiment through cooperative study or independent individual study. Generally each experimenter firstly puts up his own experiment ideas, then experiment organizer forms total experiment conception on the basis of many experiment ideas. Secondly the experiment organizer divides total experiment work into each experimenter. Thirdly each experimenter begins his divided study according to experiment division. Finally, with the help of synthetic discussion function, the experiment organizer collects all study results from each experimenter, he will sum up all study results and will form final experiment conclusions.

Antagonism and Deduction Mode: Antagonism and deduction mode is a senior mode of joint combat experiment. It is based on the above two experiment modes. In this mode, combat experiments are full of military encounters from both-sides or multi-sides. Experiment roles are divided into director side, blue side, red side, or multi-side. Generally, they begin to play a counter move according to combat plans. Their plans come from their respective collective decisions. Each move is arbitrated by white cell and they might use a simulation to assist in determining outcome of each move. An experiment might involve fighting the same campaign using two or more different strategies. After each move, they begin observing and analyzing combat situation and forming next round's combat plan and decision by collective discussion till all experiment rounds end.

Demonstration Mode: The mode isn't for pure demonstration but for improving experiment. The mode expands experiment function, means, object, domain and service scope. Demonstration experiment mode includes 3 essential experiment factors: experimenter, experiment means, experiment object, so demonstration mode is thought as a kind of expanded experiment mode.

B. Joint Combat Experiment Phrase

Joint combat experiment is divided into 3 phrases: putting up experiment ideas, arguing experiment ideas, forming experiment conclusions.

Phrase 1: Putting up experiment ideas. Main task is forming problem-tree, constructing experiment frame, forming initial experiment ideas.

Phrase 2: Arguing experiment ideas. Main task is arguing each problem in problem-tree by collective discussion, improving initial ideas and forming better plans, finally getting conclusions of each problem. Phrase 2 is a very important phrase.

Phrase 3: Forming experiment conclusions. Main task is summing up conclusions of all problems and getting total experiment conclusions.

I I I. Experiment Roles and Organization[2]

A. Experiment Roles

We have known that there are 4 modes of joint combat experiment. They are as followings: individual study mode, collective discussion mode, antagonism and deduction mode, demonstration mode. In the above-mentioned 4 modes, collective discussion mode is the most representative mode because it makes experimenter

AP-PROC-CS-09CN005

almost use all functions of joint combat experiment platform. So we select this mode as a typical example to elaborate general organizing and implementing method of joint combat experiment.

In collective discussion mode, experiment roles are divided into presenter, discusser, technical supporter.

Presenter: There is only one presenter in one times of experiment. Often only that person organizing experiment subject acts as an experiment presenter. Military expert may also act as the experiment presenter in need. Before discussion, the presenter is responsible to make experiment plans, disintegrate experiment problems, and assign work to experimenters. During discussion, the presenter guides and presents all activities of discussion according to experiment conception and frame. He is deputy for summarizing all discussion problems and putting up next step of discussion problems. After discussion, he summarizes answers of all problems and forming total experiment conclusions. At last, he must put up the next discussion plan and problems.

Discusser: In general, only military persons taking part in experiment act as discussers. The number of discussers is decided by the number of discussion problems, experimenters and domain experts. Generally the number of discussers shouldn't surpass 40 persons. But the number of auditors is not limited in collective discussion. Before discussion, work of discusser is doing enough preparation according to presenter's arrangement. He needs looking up materials, information analysis, combat calculation and puts up experiment conception. During discussion, work of discusser is introducing his own experiment conception. At the same time, he should take part in other discussers' discussion activities. After discussion, work of discusser is analyzing discussion results and getting initial discussion conclusions.

Technical Supporter: Generally technical person acts as technical supporter. Often there are 2 or 3 technical supporters in one times of experiment. Their main duties are helping experiment presenter operate computers, run software, manage experiment subjects and record discussion information.

B. Experiment Organization

According to experiment mode and experiment phrase division, one times of experiment is divided by 3 phrases: putting up experiment ideas, arguing experiment ideas, forming experiment conclusions.

(A)Putting up Experiment Ideas: Putting up experiment ideas includes 3 steps. Detailed description is as followings.

Disintegrating Study Problem, Forming Discussion Frame. According to division of experiment, every discusser puts up his problems by individual study. Then all problems are converged into a total problem frame. In the process of convergence, directories and frames of experiment subject should be list out one by one according to their classed levels. So a series of study problems will be list out under their respective directories, and an entire problem tree will be formed. When confronting with complicated problem, we should closely disintegrate the complicated problem into many small questions as simple as possible. These small simple questions may be accurately answered in a short passage of words, a choice, a group of data, a simple picture or an information form. Problem tree may be put up by presenter or his agent, or may be unrestrainedly produced by collective discussion. As soon as the problem tree is got, an initial discussion tree will be got. The initial discussion tree is also called discussion frame.

Cutting out Problem Tree. Cutting out problem tree means *tailoring* work. We need delete or add some problem branches and knots by collective discussion, expert evaluation or collective hand vote. The *tailoring* work generally is finished by experiment presenter or his agent, and it may be finished by experimenters' collective discussion or collective votes.

Forming Initial Experiment Ideas. After tailoring process, problem tree will become initial experiment ideas. Since then, all experimenters begin solving their respective problems by individual study according to division of presenter. They will do enough preparation work and obtain their answers. Some experimenters need looking up some information and materials in the lab information-lib, and accumulate enough primitive materials and knowledge. Sometimes they need doing some calculation with the lab model-lib. They need taking good use of their knowledge and get solutions to their respective problems. Their solutions would include 2 parts: Part 1 might be their viewpoints and conclusions, Part 2 might be description of their arguments. When all solutions have been gathered into total conceptions, initial experiment ideas will be formed.

(B)Arguing Experiment Ideas: Arguing experiment ideas is very important. In this phrase, initial experiment ideas of Phrase 1 will be used into Phrase 2. Implementing steps of Phrase 2 are as followings.

Collective Discussion: Generally experiment presenter presides collective discussion in experiment hall, synthetic discussion rooms or synthetic combat rooms. Often every discusser firstly introduces his distributed problems, then he elaborates his solutions and conclusions, finally domain experts begin discussing whether his solutions and conclusions are correct. Discusser may have one or more solutions and conclusions, other discussers can give their advice on improving, modifying and amending his solutions and conclusions. During the procedure of discussion, problems may be discussed one by one. When a specified problem is discussed, one discusser can give his viewpoints on the specified problem. On the other hand, discussion topics may be discussed one by one, or whole experiment subject may be discussed in one times of discussion. When discusser introduces his viewpoints, he may use all needed lab functions to prove his viewpoints. For example, he may call information looking-up function, combat calculation function, simulation and evaluation function to live demonstrate his experiment. Of course, he may also call for his own or other person's study results to assist in his viewpoints or discussion.

Statistical Analysis and Further Argument: Experiment presenter should summarize all discussers' discussion and

forms final conclusions on one specified problem. Then the presenter or his agent should carry out statistical analysis on the specified problem. The statistical analysis can display distribution of expert opinions. Type of problem may be description-type question, plot-chart-type question or combat organization-into-group-type question, etc. When confronting with the three types of questions, we should again discuss those questions which produce bigger differences of expert opinion. When confronting with statistical analysis in choice-type or data-type question, we should again discuss those questions which produce sharper deviation and less common consensus in expert answers, and so on, till we have got distinct answers to all questions. At last, we should put the viewpoints, data, argument grounds, common consensus into conclusion-area of experiment system and save them.

In the whole of argument phrase, information-support function, model-support function, military assignment-support function and discussion-support function provide collective discussion with strong support. When discusser introducing his conception and plan, he may use discussion-support function and military assignment-support function to make his demonstration more excellent in both pictures and text. When other discussers raise some doubts or suspicions, he may use information-support function and model-support function to provide many newer complement and argument. When experimenter summarizing his viewpoints, he may use discussion-support function and model-support function to make his conclusions more reasonable.

(C)Forming Experiment Conclusions: After all questions have been discussed and major common consensus has been obtained, we need sorting out and drawing total conclusions. Firstly we use result-process branch system to help us sort out conclusions of all questions. The branch system will automatically collect every question's answers and forms total conclusions, then a study report will be automatically produced. At the same time, directories and main viewpoints of the study report will be automatically produced. If necessary, we should modify the study report again through further collective discussion till we get the final version.

IV. An Example of Large Joint Combat Experiment[3]

The above analysis may be simply generalized to *4-mode and 3-phrase theory*. This theory is a basis of one times of small combat experiment. Usually large joint combat experiment is composed of many small combat experiments, so *4-mode and 3-phrase theory* will be synthetically applied to large joint combat experiment.

Large joint combat experiment is often organized in the form of experiment subjects, special experiment topics or experiment items. Namely, a large experiment subject is divided into many small special experiment topics, and small special experiment topic may be divided into smaller experiment items if need.

Now we give an example of large joint combat experiment. Assuming that a large joint experiment is a subject and the subject is divided into 5 special topics. They are: target analysis topic, combat calculation topic, combat plan study topic, simulation and evaluation topic, experiment summary topic. Taking target analysis topic as a typical example, we can elaborate its implementing method. Furthermore, we can deeply understand implementing method of the whole large joint experiment.

(A)Target Analysis Experiment Topic: Military persons are responsible for target analysis topic. They mainly study which targets would be struck. Their purpose is forming a striking target set. Specific steps are: putting up a striking target set, arguing rationality of the striking target set, forming conclusions of striking target set. Detailed descriptions are as followings.

Putting up a Striking Target Set. Implementing steps are: experimenters begin their respective study in individual study mode; presenter forms initial conceptions of target analysis after summarizing all individual results, experimenters roughly select initial striking targets in collective discussion mode; experimenters look up information of roughly selected targets and begin individual study again; experimenters carefully select striking targets by one or more times of collective discussion; after several rounds of individual study and collective discussion, presenter forms a striking target set.

Arguing Rationality of the Striking Target Set: A striking target set has been got in the previous phrase. We should prove whether the striking target set is rational. This demands that we should use quantitative analysis method and qualitative analysis method to prove its rationality. Under this condition, we need building up a target-value-evaluation system and obtain its help. With the help of the target-value-evaluation system, many field experts begin evaluating target's value, obtain statistical analysis, and get a target value sequence. On the basis of the target sequence, we can get a final target set through careful target selection and collective discussion.

Forming Conclusions of Striking Target Set: On the basis of arguing rationality of the target set, a study report on target analysis topic will be formed. The report's name is *Study Report of Target Analysis Topic*.

In target analysis experiment topic, experiment method of each phrase isn't quite the same. The purpose of phrase 1 is putting an Initial striking target set. So a quantitative analysis method is mainly used in phrase 1. The purpose of phrase 2 is arguing the striking target set. So integrating method is used in phrase 2. The integrating method is namely integration method of quantitative analysis method and qualitative analysis method. We can see, individual study mode and collective discussion mode are alternatively used in phrase 1. For example, firstly use individual study, secondly use collective discussion, thirdly use individual study again, fourthly use more collective discussion, and so on. We can also see, *4-mode and 3-phrase theory* is synthetically applied to target analysis topic.

In a word, target analysis experiment topic not only embodies synthetic application of experiment theory, but also reflects total method of joint combat experiment.

Although there are some differences among 5 different experiment topics, their experiment method is basically similar. In order to fully see total method of joint combat experiment, the following parts will give a brief description on other topics.

(B)Combat Calculation Topic: Combat calculation topic is responded by military persons and technical persons. Military persons are responsible for forming striking plan of targets, and technical persons are responsible for combat calculation which includes combat calculation of armed services and all branches.

This topic mainly uses qualitative calculation method and collective discussion method. Its specific implementing steps are: determinate demands of damage; sort out important order of striking targets; select striking weapons; analyze striking effects(include rough calculation and accurate calculation); optimize weapons selection.

When these steps have been finished, a striking plan will be formed and combat calculation report will be got.

(C)Plan-Studying Topic: The purpose of plan-studying topic is sufficiently arguing a firepower striking plan and forming a combat plan. During the argument course, military persons and technical persons jointly determinate numbers of forces and weapons according to results of combat calculation, results of target analysis as well as military postures of both-sides or multi-sides. They not only determinate combat resolve, but also make combat plan.

With the help of function of information-support, model-support, military assignment-support function and discussion-support function, they will further their argument on rationality and closeness of firepower striking plan.

(D)Simulation and Evaluation Topic: Simulation and evaluation topic is responded by technical persons. When technical persons have input all parameters into combat models, software systems begin running computer-deduction, combat-calculation, statistic analysis, data record. At last, software systems begin saving combat effects. All these works are automatically finished and whole process doesn't need person intervention.

(E)Experiment Summary: When all experiment topics have been finished, the experiment presenter will sum up all experiment results. Its steps are: replay experiment course and extract experiment results; synthetically analyze experiment data and form total conclusions; write an experiment report; submit, censor and save the final experiment report.

REFERENCES

[1] Preparation group of the 262th Xiangshan science meeting. Theory and application on hall for workshop of metasynthetic engineering. Beijing: Xiangshan science meeting, 2005

[2] Han Yan-fei, Jiang Jing-zhuo.Analysis on development of metasynthesis theory and technology[J]. Military operations research and systems engineering, 2006,20(1):3-7

[3] Zhao Cun-ru, Li Ning, Wang Wei. Complexity of war and military systems engineering[J].Military operations research and systems engineering, 2006,20(4):70-73

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 475-478

Advanced OFDM System for Modern Communication Networks

Pingxiang Yao
Department of Navagation Qingdao Ocean Shipping Mariners College, Qingdao, China, 266071
Email: sdclypx@yahoo.com.cn

Abstract— **Orthogonal frequency division multiplexing (OFDM) has become one of the most important modulation methods in many fields, such as high-speed communication systems. It is proposed that OFDM technique can offer variable bandwidth, improved protection to shadow and multipath fading and enhanced robustness thanks to the insertion of the guard interval. In this paper, we introduced the principles of OFDM at first, then analysised the advantage and disadvantages of OFDM system, and finally discussed its application in contemporary high-speed communication systems, especially in wireless mobile networks.**

Index Terms— **OFDM; ISI; GI; IDFT; DFT**

I. INTRODUCTION

With the development of wireless communication technologies and the emergence of a large number of multimedia services, the speed and reliability of data transmission are expected to be increased more higher. Orthogonal frequency division multiplexing (OFDM) system is proposed that it can offer variable bandwidth, improved protection to shadow and multipath fading and enhanced robustness thanks to the insertion of the guard interval. So from a certain extent, the OFDM system can solve the bandwidth requirements for data user [1]. And OFDM modulation for wireless network technology is also designed to more fully utilize existing bandwidth, and can very well against the frequency selective fading and narrowband interference. Therefore, using OFDM technique to provide high-speed data transmission has become a hot topic in wireless communication field.

II. PRINCIPLES OF OFDM

OFDM is a multicarrier transmission technique which divides the available spectrum into many subcarriers, where each subchannel is modulated by a low rate data stream. These subcarriers are made orthogonal to one another, that is, independent from each other. The orthogonality of the subchannels means that each subcarrier has an integer number of cycles over a symbol period. Owing to this, the spectrum of each subcarrier has a null at the centre frequency for the other subcarriers in the system. This results in no interference between the subcarriers, allowing then to be spaced as closed as theoretically possible, i.e. their spectrum can be overlapped, as shown as figure 1. Compared to conventional frequency division multiplexing (FDM) system, the spectrum efficiency is improved greatly in OFDM system [2].

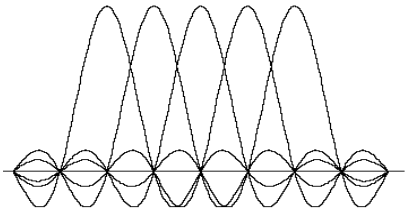


Figure 1 OFDM spectrum

While there are N subcarries in a OFDM symbol, in which each subcarrier adopts phase shift keying (PSK) or quadrature amplitude modulation (QAM) and carries the data symbol di(i=0,1,2,……N-1), and T is the duration of a single OFDM symbol, fi is the frequency of the ith subcarrier, the rectangular function rect(t)=1, $|t| \le T/2$,Then the OFDM symbol s(t), which beginning at t=ty, can be written as [1]:

$$s(t) = \mathrm{Re}\left\{\sum_{i=0}^{N-1} d_i rect(t - t_s - T/2)\exp[j2\pi f_i(t - t_s)]\right\} \qquad t \le t_s + T$$
$$s(t) = 0 \qquad t \prec t_s \text{ or } t \succ T + t_s \tag{1}$$

Before assigned into the various subcarriers, the input high-speed serial data is mapped onto the amplitude and phase components after FSK / QAM modulation. Usually the equivalent base-band signal which used to describe the output of OFDM signal can be expressed as following:

$$s(t) = \sum_{i=0}^{N-1} d_i rect(t - t_s - T/2)\exp\left[j2\pi \frac{i}{T}(t - t_s)\right] \qquad t_s \le t \le t_s + T$$
$$s(t) = 0 \qquad t \prec t_s \text{ Or } t \succ T + t_s \tag{2}$$

Which, the real and empty parts of s(t) correspond to the same OFDM phase and quadrature values respectively [1]. That is multiplied with the sinusoid components of every subcarriers separately in the actual system, and constitutes the final subchannel signals and the synthesis OFDM symbols. The basic structure of OFDM system is given as Figure 2, in which fi = fc + i / T. The subcarriers demodulation is completed at the receiving end, that is, the same phase and quadrature vector will be transformed to time-domain data to build up the original data and information.

AP-PROC-CS-09CN005

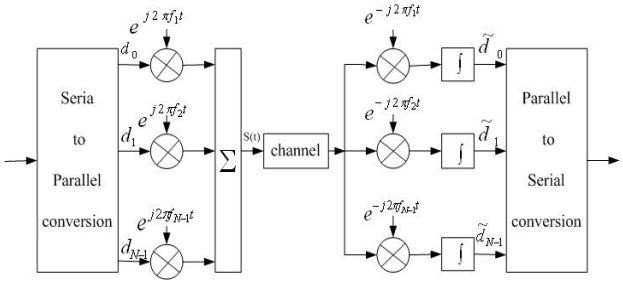


Figure 2 Basic Structure of OFDM System

In communication systems, the analog implementation of OFDM can be extended to the digital domain by using the discrete Fourier Transform (DFT) and its counterpart, the inverse discrete Fourier Transform (IDFT). These mathematical operations are widely used for transforming data between the time-domain and frequency-domain. These transforms are interesting from the OFDM perspective because they can be viewed as mapping data onto orthogonal subcarriers. For example, the IDFT is used at the transmitter to modulate the original OFDM data, which is to take in frequency-domain data and convert it to time-domain data. In order to perform that operation, the IDFT correlates the frequency-domain input data with its orthogonal basis functions, which are sinusoids at certain frequencies. This correlation is equivalent to mapping the input data onto the sinusoidal basis functions. And DFT can be used at the receiving end to utilize the demodulation.

In practice, OFDM systems are implemented using a combination of fast Fourier Transform (FFT) and inverse fast Fourier Transform (IFFT) blocks that are mathematically equivalent versions of the DFT and IDFT, respectively, but more efficient to implement. An OFDM system treats the source symbols (e.g., the QPSK or QAM symbols) at the transmitter as though they are in the frequency-domain [3]. These symbols are used as the inputs to an IFFT block that brings the signal into the time-domain.

According to figure 2, if we order s(t) = 0 in formula (2), ignore the rectangular function, and take the sampling rate of signal s(t) at T/N, where N is the number of subcarriers and T is the IDFT input symbol period mentioned above, then we can attain that t=kT/N， (k=0,1,…,N-1). Thus equation (2) can be condensed as:

$$s_k = s(kT/N) = \sum_{i=0}^{N-1} d_i \exp\left(j\frac{2\pi ik}{N} \right) \qquad (3)$$

$$0 \le k \le N-1$$

Through the N-point IDFT calculation, frequency-domain data symbol di becomes time-domain data symbols sk, and it will be sent into the wireless channel after radio frequency (RF) modulation. Each IDFT output data symbols sk is the summation of all N subcarrier signals, thus, the IDFT block provides a simple way to modulate data onto N orthogonal subcarriers. The block of N output samples from the IDFT make up a single OFDM symbol. The length of the OFDM symbol is NT where T is the IDFT input symbol period mentioned above.

After some additional processing, the time-domain signal sk is transmitted across the channel. At the receiver, a DFT block is used to process the received signal and bring it into the frequency-domain. Ideally, the DFT output will be the original data symbol di that were sent to the IDFT at the transmitter:

$$d_i = \sum_{k=0}^{N-1} s_k \exp\left(-j\frac{2\pi ik}{N} \right) \qquad (0 \le i \le N-1) \qquad (4)$$

Since the high-speed serial input data stream is assigned into N parallel subchannels after a serial/parallel conversion in OFDM system, the symbol rate in each subcarry is greatly reduced, and the impact of multipath is weakened evidently. Therefore, the OFDM system is resistant to multipath delay spread. In order to eliminate the inter-symbol interference (ISI) to the minimum, the guard interval (GI) is introduced between the OFDM symbols [2]. The GI length of the guard interval Tg should be longer than the largest channel delay spread so that the multipath component of such symbol will not interfere with the next one. The OFDM symbols with Guard Interval are shown as Figure 3.

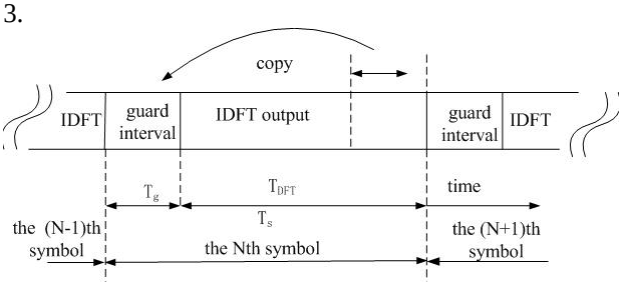


Figure 3 OFDM Symbols with Guard Interval

In the figure 3 above, the total length of each symbol is TS = Tg + TDFT, where TS is the total length of a single OFDM symbol, Tg is the length of guard interval, TDFT the length of OFDM symbol generated by DFT transformation without guard interval. Then at the receiving end, the sampling start time Tx should meet the following formula:

$$\tau_{\max} < \mathrm{TX} < \mathrm{Tg} \qquad (5)$$

In formula (5), $\tau_{\max}$ is the maximal multipath time span of each channel. When the sampling start time Tx satisfies formula (5), the disturbance, caused by previous marks, will only exist in the duration of $[0, \tau_{\max}]$. When there are more subcarriers, the duration of OFDM symbols Ts is longer than channel's pulse response length. For this reason, the ISI influence is very small. If the guard interval lasting time Tg conforms to Tg≧ $\tau_{\max}$, the shortage of ISI may be overcome completely. At the same time, the number of subcarriers' period which existing in the OFDM delay transcription is integer, the delayed signal will generates inter-channel interference （ICI）during the demodulation. Figure 4 is shown as the

OFDM structure, which is based on IDFT/DFT and inserted guard interval.

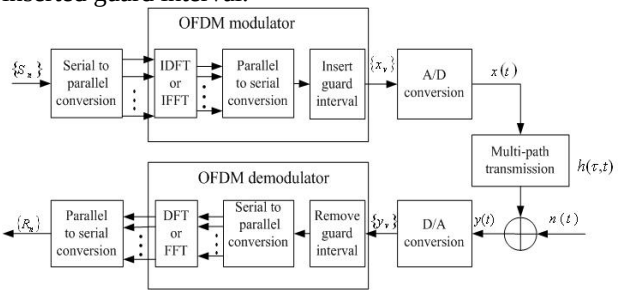


Figure 4 IDFT/DFT-Based OFDM Systems

In figure 3, Tx is the duration of OFDM symbol, and it meets the following equation:

$$T_S=T_g+T_{FFT} \quad (6)$$

The separate length of guard interval, namely the number of sampling point is:

$$L_g \geq \left\lceil \frac{\tau_{\max} N}{T_s} \right\rceil \quad (7)$$

And according to Figure 3, OFDM sampling sequence $\{x_v\}$, which containing the guard interval and being power- normalized, can be expressed as:

$$x_v = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} S_n e^{j2\pi nv/N}, \quad v = -L_g, \cdots, N-1 \quad (8)$$

After traveling through wireless channel, the receiver signal $y(t)$ is mixed with channel function $h(t)$ and Gaussian white noise n(t). It can be written as:

$$y(t) = \int_0^{\tau_{\max}} x(t-\tau)h(t,\tau)d\tau + n(t) \quad (9)$$

After A/D conversion, the received signal $y(t)$ is changed to sequence $\{y_v\}$, $v = -L_g, \cdots, N-1$, which is the digital sample of $y(t)$ at the rate of T/N. ISI will only interferes with the Lg preceding sampling points among the received sequences $\{y_v\}$. Therefore, the influence of ISI may be eliminated absolutely after removing the Lg sampling points. And the sequence $\{y_v\}$, $v = -L_g, \cdots, N-1$, which deletes the guard interval, will be carried on the DFT transformation and generated the output multi-carrier demodulation sequence $\{y_v\}$ with N complex points:

$$R_n = \frac{1}{\sqrt{N}} \sum_{v=0}^{N-1} y_v e^{-j2\pi nv/N}, \quad n = 0, \cdots, N-1 \quad (10)$$

If the number of subcarriers N is chosen appropriately, it may make the channel influence flatten. And the insertion of guard interval helps to fortify the orthogonality between subcarriers also. So the influence of ISI and the ICI generated by multipath environments can probably be removed totally in OFDM system [4].

The received signal in frequency domain can be displayed as following:

$$R_n=H_nS_n + N_n, \quad n=0,\ldots,N-1 \quad (11)$$

Among them, Hn is the complex fading coefficient of the nth subcarrier, Nn represents the additive Gaussian white noise of the nth subchannel, and its real and empty parts separate from each other while both following zero mean value Gaussian distribution .The noise variance is:

$$\sigma^2 = E\left\{ \left|N_n\right|^2 \right\} \qquad n = 0, \cdots, N-1 \quad (12)$$

III. PROPERTIES OF OFDM

OFDM is a modulation scheme that has recently gained immense popularity in the design of wireless communication systems. The main advantages of OFDM over other communication techniques are that:

Makes efficient use of the spectrum by allowing overlap.

By dividing the channel into narrowband flat fading subchannels, OFDM is more resistant to frequency selective fading than single carrier systems are.

It solves the problem of inter symbol interference (ISI)

Using adequate channel coding and interleaving one can recover symbols lost due to the frequency selectivity of the channel.

Channel equalization becomes simpler than by using adaptive equalization techniques with single carrier systems.

It is possible to use maximum likelihood decoding with reasonable complexity.

OFDM is computationally efficient by using FFT techniques to implement the modulation and demodulation functions.

Is less sensitive to sample timing offsets than single carrier systems are.

Provides good protection against cochannel interference and impulsive parasitic noise.

Different from single carrier system, the OFDM output signal at receiving end is the summation of which transferred in a great large number of orthogonal subchannels. Thus, there are also disadvantages for OFDM technology:

The OFDM signal has a noise like amplitude with a very large dynamic range; therefore it requires RF power amplifiers with a high peak to average power ratio.

It is more sensitive to carrier frequency offset and drift than single carrier systems due to leakage of the DFT.

IV. APPLICATIONS IN MODERN COMMUNICATION NETWORKS

OFDM is a transmission technique used to achieve very high data rates. It is applied in several contemporary communication systems, including asynchronous digital subscriber line (ADSL), digital audio/video broadcast systems, wireless area network (WLAN), broadband wireless access (BWA), as well as 3G CDMA or 4G cellular mobile networks [5].

In 1997, OFDM was first brought into operation in digital audio broadcast (DAB) standards, which raised by

European telecommunications standards institute (ETSI) [6]. Nowadays, it has been popularly used as the modulation process in radio broadcasting systems, such as digital audio broadcasting (DAB), digital video broadcasting (DVB), as well as high definition television (HDTV).

In WLAN field, both the U.S. IEEE 802.11a and European ETSI Hiperlan/2 standards utilize OFDM technology [5]. IEEE802.11a works in 5 GHz frequency region and employs OFDM modulation in its physical layer. European ETSI Hiperlan/2 standards utilizes both OFDM and link autoadaptation technologies in physical layer, and adopts connection-oriented TDMA/TDD and asynchronous transfer mode (ATM) methods for media access control (MAC) Layer. The maximum data rates can achieve 54Mbit/s. OFDM works well in home and office environments for handling wall reflections and movement within the structure.

OFDM is also suitable for BWA because of its super properties. IEEE 802.16 work group is responsible for the technological works of BWA. It has already developed a new standard for BWA - IEEE 802.16a, which works at 2GHz-11GHz and adopts OFDM technique at physical layer. As a new wireless access technology, IEEE 802.16a may also promote the future development of cellular mobile communication networks.

At present, the 3rd generation (3G) cellular mobile networks based on code division multiple access (CDMA) have been built up and put into service. But the maximum data transmission rate provided by 3G is only 2Mbps, and is lower in practical system. So 3G network do not suit the demands of data user for multimedia services. In order to promote the 3G properties, OFDM is introduced into CDMA systems, and the combination of CDMA and OFDM is being investigated by many researchers in mobile field [7].

Nowadays, frequency has already be scarce more and more as the resources of mobile communication systems because of the enlargement of network scale, the increscent of user's requirement for service quantities, as well as the high data transmission rates. So in the 4th generation mobile cellular network (4G), OFDM will get more extensive application because of its fine performance.

V. CONCLUSIONS

The multimedia services become one of the developing directions in the wireless communication systems. The multimedia service demands high-speed data transmission rate. Therefore, those techniques which supporting high-speed data transmission are the inevitable development trend in wireless communication networks. OFDM is such a technology that can offer high-efficient and reliable data transmission that has caused much more attentions in modern communication region. With the scale growth of communication networks and the user's demand for datumization, broadband, individualization and mobility, the OFDM technique will be widely used in both cable communication networks and wireless systems.

REFERENCES

[1] Bingham, J. A. C., Multicarrier Modulation for Data Transmission: An idea whose time has come, IEEE Communications Magazine, Vol.28,no.5, pp.5-14, May 1990.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Alen, T. C. H., A. S. Madhukumar, et al. (2003), "Capacity enhancement of a multi-user OFDM system using dynamic frequency allocation." Broadcasting, IEEE Transactions on 49(4): pp: 344-353.

[3] Rhee, W. and J. M. Cioffi. Increase in capacity of multiuser OFDM system using dynamic subchannel allocation, in Proceedings of Vehicular Technology Conference Proceedings, 2000, VTC 2000-Spring Tokyo, 2000 IEEE 51st, 2000. 2: pp: 1085-1089 vol.1082.

[4] J. Stott, The Effects of Phase Noise in COFDM, EBU Technical Review, Summer 1998.

[5] V. Erceg et al., Channel models for fixed wireless applications IEEE 802.16 Broadband Wireless Working Group, IEEE 802.16a-03/01, Tech. Rep., 2003.

[6] ETSI DVB-T, "EN 300 744 v1.5.1 Digital Video Broadcasting (DVB) Framing Structure, Channel Coding and Modulation for Digital Terrestrial Television, " Tech. Rep., June 2004.

[7] H. Sampath, S. Talwar, J. Tellado,V. Erceg, and A. Paulraj, "A fourth-generation MIMO-OFDM broadband wireless system:Design, performance and field trial results," IEEE.Commun. Mag, vol.40,no.9,pp.143-149, Sep. 2002.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 479-482

Study on Circumvention Measures of Credit Information Security Risks in E-Commerce

Xiaoming Meng
School of information, Guangdong University of Business Studies
Guangzhou, Guangdong, China
mxm_me@163.com

***Abstract*—Brief introduced the concepts of credit and credit risk in e-commerce, analyzed the impact of credit risks in e-commerce. Gave out the basic contents of personal credit information and business credit information in e-commerce, and analyzed and pointed out the main factors that caused the security risks of e-commerce credit information. At last, proposed some technical and non-technical measures to circumvent the credit information security risks in e-commerce, such as to improve the infrastructure of E-commerce and to strengthen researching on information security technology, to establish a unified platform for e-commerce transactions, to achieve the integration and sharing of the credit information resources, to obey the security E-commerce operation rules and to train good habits of online transaction, to establish and improve the social credit system, to strengthen the laws of being built, operated, to establish the system of punishing dishonesty and invigorating mechanism, to strengthen and to improve the credit information management, to regulate credit information services, and to train and improve the credit information security awareness.**

***Index terms*—e-commerce, credit, credit information, security risks**

Ⅰ. INTRODUCTION

E-commerce has gradually become the dominant mode of modern business, and it has virtual and non-contact characteristics. In E-commerce, because the trade transactions between the two sides have been took by network , so both of tem can not intuitively understand each other and can not seen the transaction subjects, they can only make trading decisions by the information obtained online and subjective judging. However, due to false trading, counterfeiting, fraudulent contracts, credit distortion, bid up the value of subject by online auction etc., violates the legitimate rights and interest of consumers, and other illegal activities often occur, all these lead to more and more consumers will loss of confidence in E-commerce, and make them are not willing to choose E-commerce. There are many reasons make consumer lack of trust in E-commerce. But we consider that lack of credit has become a major obstacle during E-commerce developing. Therefore, an effective solution the credit problem in E-commerce and to foster consumer trust in E-commerce has become the key to develop E-commerce. The primary task to effectively solve the credit problem in E-commerce means to make the credit information security for the main body of transaction. So, to study and discuss the security risks of E-commerce credit information has important practical significance for building and improving E-commerce credit system.

Ⅱ. CONTENT OF E-COMMERCE CREDIT INFORMATION

E-commerce credit information refers to the records which are created by enterprises and individual consumers in E-commerce activities, and relates to the credit behaviors, as well as the worthy information for using to evaluate their credit level. Accord to the different subjects of credit, the credit information in E-commerce can be divided into enterprise's credit information and personal credit information [1].

A. Enterprise's Credit Information

Enterprise's credit information refers to the records and the relation data which formed in credit transactions to fulfill or non-fulfill obligations. It Includes:

(1) Basic information. It includes registration data (for examples the enterprise's name, registered address, legal represent, registered capital, type, business scope, setup date, business period etc.), relationship information, rights of import and export, financial information etc.

(2) Situation information of affiliates, subsidiaries, shareholding enterprises.

(3) Information of capital construction. It includes shareholders, investment ratio, the rate of availability capital, investment methods etc.

(4) Information of Board of Directors and managers. It includes board charters, organization structure etc.

(5) Information of production and business status. It includes production-marketing situation, costs-earnings, production technology, production efficiency etc.

(6) Information on the financial status. It includes asset situation, profit situation, account receivable/payable etc.

(7) Exchanges bank information. It includes loan bank, loan amount, loan reputation and so on.

(8) Litigation records. It includes the records of civil, administration, legal etc.

B. Personal Credit Information

Personal credit information is the records and relative data which are created by individual who performance or non-compliance in credit transactions. For the content

and scope of personal credit information, there are different requirements in different countries. We think that it should include:

(1) Identifying information. It includes name, sex, original place, date of birthing, identity card number, occupation, work and education experience, political orientation, home address, marital status, spouse information, health status, telephone, E-mail address etc.

(2) Business Information. It includes personal income, assets, credit records, records of public utilities service, as well as individuals actual performance records which relate with finance institutions or housing fund management center and other organs on loans, guarantees, credit cards, insurance etc.

(3) Public information. It includes personal taxes, participate in social insurance, and personal property status and changing etc.

(4) Specific records. It includes the honorary title, awards, and records which may affect the personal credit situation, such as civil, criminal, administrative proceedings and administrative punishment.

(5) Other information. It includes the credit-related information in addition to the above range.

III. CAUSING REASONS OF CREDIT INFORMATION SECURITY RISKS IN E-COMMERCE

There are many reasons to cause credit information security risks in E-commerce, the main reasons are:

A. Defects of Techniques in E-commerce

(1) Because the network has opening characteristics, so it gives some illegal people or organization a choice to steal the user's credit information.

(2) The reliability of network is questioned. Sometimes it will be interrupt by some reasons.

(3) The security of network is questioned. The malicious attacks, illegal invading, sniffers, monitors and other illegal behaviors will produce security risks.

(4) The database system may be existing unsafe linking or backdoor for attackers.

(5) The defects of transacting software and bad habit of developing application program will remain "backdoor" for cracks.

(6) The bad habits of online transactions will give a choice to attackers.

(7) The number system is not standard and uniform will cause credit information security risks.

B. Characteristics of E-commerce Transaction

(1) Information of each side in transaction is not asymmetric. This will cause credit information security risks.

(2) Property boundary of E-commerce transactions will be fuzzy. It will make the both sides of transaction difficult to sign equality contract.

(3) Game between the main bodies of transactions [2]. In E-commerce, the sale is always opposing with another on interest, so each one will be in order to his own interest, may deliberately modify his credit information and provide false credit information to another in order to win the trust and to achieve his own commercial purpose.

C. Driven by Market Profits

Now, the data warehouse technology and data mining technology are developing quickly. Some illegal individuals or organizations use them to process and use user's credit information in E-commerce to obtain reap benefits, and trigger credit information security risks [3].

D. Inadequate Laws and Regulations

(1) Inadequate laws and regulations. Because of E-commerce started late in China, the construction of corresponding laws and regulations delays to the E-commerce rapid development. This situation leads to the relative laws and regulations have still existed many shortages, although there are many items increased in our in the Constitution law, civil law and many other laws and regulations. Such as "electronic signature law" has been published two years ago. But the construction of laws on business information protection and privacy protection is still not perfect. Particularly, the laws of security protection on the credit information of main body in E-commerce are still to be strengthened.

(2) The intensity of legal punishment for dishonesty is not very strong. Although it has a certain constraint for "dishonesty" and "false" credit phenomenon in some laws and regulations. But the intensity of punishing and enforcement of performing are still not strong enough, and the effectiveness of restraining is still less than ideal.

(3) A lack of credit incentive mechanism. The incentives mechanism for honesty and trustworthiness of the main body in E-commerce is still underdeveloped.

E. Imperfect of Credit Information Management and Service

(1) Imperfect of social credit information management system. At present, the social credit system is still imperfect, and the large database network of social credit information which covers the whole society is still setup. It leads to that does not realize "anytime, anywhere" querying the credit information of main body in E-commerce. In addition, some businesses or individuals ignore the role of credit information management, and even, can not understand the credit information management. They think that this may affect the relationship with customers, and does not adopt effective measures to manage and organize the credit information. And, lacking of professionals is still an important reason for poor management.

(2) Low level of credit management and credit information intermediary service. The so-called credit is a credit agency which uses various means to extensively collect, process credit information to verify the credit status of the survey object. However, the level of some credit management and credit intermediary service organizations is not high, and the security management of credit information is poor, and the credit information is incomplete, data updated is not timely and other factors,

leads to the result of credit management is distortion, and leads to the security risks of credit information.

F. Human Factors

(1) Credit awareness is not high. In the current market environment, the "supreme interest" dominates the ideology of some ones in business and their sense of "honest, trustworthy, fair competition" is weak, and they neglect the credit information and its security protection.

(2) Credit ethical standard is not high. Credit belongs to the realm of ethics. However, the credit is extremely pale in some ones' moral thought. They lack of good moral cultivation, and wantonly spread, distort other's credit information to fulfill their own interest.

(3) Impacting of the traditional economic system. In the traditional economic system, the two sides of transactions are face to face. Their credit status is evaluated by past "dealings", so the credit has a profound sense. But in E-commerce, the credit situation is determined by analyzing and evaluating the credit information on all aspects of the main bodies in transactions. Therefore, the traditional economic system make the main bodies of E-commerce weaken their emphasis extent on credit information, and this will indirectly cause credit information security risks.

Ⅳ. CREDIT INFORMATION SECURITY RISKS AFFECTING TO E-COMMERCE

In E-commerce, the credit information security risks have a strong influence either for main bodies of transaction or for E-commerce Website. It shows bellow.

(1) For customers. The first is that it will affect the customers' credit level and their credibility status. The second is that it will affect to the customers' online consumption behaviors. The third is that it will cause the customer's credit information disclosed, such as consumption records, bank card number etc.

(2) For enterprises. The first is that it will affect the enterprises' credit level and their credibility status. The second is that it will affect to the direction of enterprises' E-commerce Websites. The third is that it will affect to the enterprises' privacy information disclosed, such as financial status, status of contract fulfillment, bank account etc.

Ⅴ. CIRCUMVENTION MEASURES OF CREDIT INFORMATION SECURITY RISKS IN E-COMMERCE

A. Technical Measures

(1) To improve the infrastructure of E-commerce and to strengthen researching on information security technology

The first is that the supporting technology system of E-commerce should be improved continuously, Such as authentication, signature, network and information security, payment, database technologies and so on. The second is that research on the technological innovation and application of information security technology should strengthen, such as improvement study on encryption, digital certificates, digital signature, intrusion detection, firewalls and other infrastructure security technology, and the application research on P3P, digital watermarking, information hiding techniques in E-commerce credit information security, and comprehensive integration research of variety security technologies.

(2) To establish a unified platform for e-commerce transactions, to achieve the integration and sharing of the credit information resources

It exits a serious bottleneck of credit information in the process of E-commerce credit system construction, for example, the credit information held by the various departments whose credit data are closed, credit information resources fragmented, and this make a great obstacle for E-commerce credit market developing [4]. Therefore, we should play the role of the government promotion to allocate special funds, lean on government to establish a credit information integration management system on enterprises and individuals by coordinating banks, business administration, public security, taxation departments etc. to achieve the integration of credit information resources and fully open, to achieve the credit information interconnection which crosses departments, crosses industries and crosses areas, to achieve query by network in the whole nation, to effectively overcome the information security risks due to information asymmetry [5].

(3) To obey the security E-commerce operation rules and to train good habits of online transaction

In E-commerce activities, to obey standard and security E-commerce operation rules is useful for improving the security of main body's credit information, and reducing the credit information security risks due to make unsafe transaction operation. Specific measures are [6]: The first is proper use and set up Cookie to avoid information disclosure. The second is to avoid using simple passwords and to reduce the risks of being cracked. The third is the ActiveX should be masked in E-commerce activities. The forth is to erase the traces of computer timely. The fifth is to implement hiding or encrypting for credit and privacy information. The sixth is to refuse being visited by threatened Websites. The seventh is to install firewall software to shield the illegal invasion and malicious attacks. The eighth is to use proxy server and to amend windows' BUG timely, to prevent the intrusion using IP fraud. The ninth is to clear the "saved files" and the log files, as well as the files accessed in favorites and temporary files folder, to avoid disclosing privacy due to these files being accessed. The tenth is to login internet by VPN.

B. Non-technical Measuers

(1) To establish and improve the social credit system. At present, there are four typical credit models adopted in E-commerce: intermediary model, guarantor model, Web site business model and delegation authority model. But whatever which model is adopted, the effective E-commerce is needed. Therefore, we must build an effective E-commerce credit system to prevent the risks. The specific measures are: According to the specific

E-commerce situation in China to cooperate multi-parties, and to make government, associations, enterprises, intermediaries developing their own credit system simultaneously, and then to be integrated, to gradually construct the social credit system at different levels, and which is interoperable, sharing, all-round, three-dimensional and efficient.

(2) To strengthen the laws of being built, operated, to establish the system of punishing dishonesty and invigorating mechanism. To strengthen building laws on credit information, to speed up the process of building the laws which relate to E-commerce credit information, such as credit, electronic transactions, electronic information protection, personal data protection, trust management, credit information agency management, and enterprises' credit information management and consumers' credit information management, etc. Having built the relation laws to make the content of E-commerce credit information, and its management responsibilities and the use of permissions being clear, to constraint and regulate the individuals or organizations querying and using and managing the E-commerce credit information, to effectively curb the illegal behaviors occurring by wanton spreading, tampering credit information and forging unreal credit information. Especially, in enforcing laws, it is important to establish mechanism for disciplining dishonesty and encouraging trustworthy, and to publish severely the legal and natural persons who have bad credit records, but to inspire that of having good credit records. In addition, it should build electronic transactions, electronic payment systems, credit card system according to the characteristics of E-commerce environment and transactions. Such as "Electronic Signature Law" which has been formally implemented in April 2005, it has played an active role to effectively protect the interests of both sides, to eliminate credit crisis, to reduce credit information risks. It has effectively improved the legal environment of China's E-commerce and reduced the credit information security risks.

(3) To strengthen and to improve the credit information management. First is to build inter-regional, cross-sectoral and inter-departmental unified credit information management system which is promoted and coordinated by the specific credit information management department of government. Second is that the special department of government should strengthen supervising for the credit agency. Third is that enterprises or individuals should fulfill reporting and updating of the obligation for their own credit information.

(4) To regulate credit information services. The services, relating to credit information, which the credit information service agencies can provide include [7]: providing credit information, evaluating credit level and using of enterprise's credit resources. In order to ensure the authenticity of the services provided to reduce the security risks of credit information, we consider that they should have a good credit level and regulate the credit information services by establishing appropriate service standard, management system, disciplinary and incentive mechanisms and so on.

(5) To train and improve the credit information security awareness of the main bodies in E-commerce. This is most important in the variety measures of preventing credit information risks in E-commerce. The content of training includes the education on credit morality and information technology.

Ⅵ. Conclusions

To sum up above, in E-commerce developing process, in order to resolve the issue of credit information security risks, and to create an honest, trustworthy, and honest credit environment for the E-commerce developing rapidly. The first is that we should to strengthen the application research on information security technologies using in credit information protection, and to develop good habit of safe online transactions, to build platforms of credit information management and its services which covers the entire society and fully open, and to effective resolve the information asymmetric problems between both sides of transaction in E-commerce. The second is that we should to accelerate and improve social credit system, to strengthen legal construction, to strict enforcement of the laws, to build the mechanism for publishing dishonesty and encourage trustworthy, and to strengthen credit information management, to standardize credit information services, to establish a correct and healthy credit ethic, and to enhance the people's security consciousness.

Ⅶ. ACKNOWLEDGMENT

This work is supported by Philosophy and Social Sciences Foundation (07O05) of Guangdong Province, China.

REFERENCES

[1] Jiping Lei. "Personal credit system and personal information protection". *National Judges College Law Journal, Peking, China*, 2006(1-2), pp.76-79.

[2] Yunxi Cheng. "Credit Risk in E-commerce". *Commercial Research, Peking, China*, 2007(11), pp.173-174.

[3] Jingwei Zhang. "E-Commerce security measures for credit". *China Storage & Transport, Tianjin, China*, 2007(3), pp.92-94.

[4] Dan Chen, Shukuan Zhao, Shunlong Gong. "Study on building the credit management system of electronic". *Information Science, Changchun, China*, 2006(1), pp.49-53.

[5] Ming Hu. "Discuss on credit problem in E-commerce". *Legal System and Society, Kunming, China*, 2007(2), pp.304-304.

[6] Xiaoming Meng "Study on the safeguard problems of public privacy information in E-commerce era". *Journal of Information(Supplement), Xian, China*, May, 2006.

[7] Qiang Bi, Zhi Qi, Yunfeng Bai. "Research on the model of credit information service of E-commerce.". *Information Science*, 2007(11), *Changchun, China*, pp.1634-1639.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 483-487

Study on Protection Measures of People's Information Privacy right in E-commerce

Xiaoming Meng
School of information, Guangdong University of Business Studies
Guangzhou, Guangdong, China
mxm_me@163.com

***Abstract*—Define the basic content of people's privacy information and information privacy right in E-commerce, analyze the creating reasons of the problem about people's information privacy right in E-commerce, point out some protection measures of people's information privacy right in E-commerce in four aspects, such as to improve people's safety awareness, to make good external environment of E-commerce, to improve the safety performance of E-commerce system, to prevent malicious attacks, and on the points of view such as the information security management, privacy information protection technology, law, management, humanities and social sciences and many other subjects.**

***Index terms*—E-commerce, privacy information, privacy right, information security**

Ⅰ. INTODUNTION

Now, the network and E-commerce is being applied widely and deeply, but at the same time, some illegal businesses or individuals driven by interests collect, theft, eavesdrop, intercept, destruct and spread the business privacy information of the people who take part in the E-commerce by adopting a variety of technical measures, and using the opening inherent weakness and technical loopholes of internet. All above invades the rights of people's information privacy, affects the people's trust for E-commerce, and constrains the rapid developing of E-commerce, and is not conducive to social stability and harmony. How does protect the people's information privacy right in E-commerce has become a focus problem that should be studied and discussed.

Ⅱ. CONTENT OF PEOPLE'S PRIVACY INFORMATION AND PRIVACY RIGHT IN E-COMMERCE

A. People's Privacy Information in E-commerce

The people's privacy information in E-commerce means that the information and data which is not known willingly in E-commerce activities, that is the people's privacy information. According to the characteristics, the people's privacy information can be divided into two categories: static privacy information and dynamic privacy information.

(1) The static privacy information means that the privacy information which is not on line, express and access in static type, such as the personal basic information, trait information, special information and trust information etc..

The personal basic information includes name, sex, age, hometown, political landscape, workplace, telephone number, address, E-mail which are used to describe the people's basic status, and job title, ID number which are used to describe the people's ID.

The trait information includes personality, voice, history, marriage and fertility status, family members, social relatives, height, weight, experience, property status, living preferences, measurements of women, etc.

The special information includes archival material, live or work or business activity logs, technical or commercial important documents, and the privacy information which is stored in a bank or certification center, computer cache, cookie, a temporary files etc.

The trust information includes loan records and the amount, business reputation, status to fulfill contractual responsibilities of law, status to implement the obligations etc. It is used to evaluate trust level.

(2) The dynamic privacy information means that the privacy information which is online, express and access in dynamic type. It includes online transacting information, secret information being transmitted online, and trace of network activities.

The online transacting information includes the account and password of credit card, consumer card and network card etc., and the information during payment and being transmitted and accessed online.

The secret information of all kinds of files being transmitted online includes e-mail, contract documents, business confidential documents, corporate strategic plan, decision programs, private documents, etc., which are transmitted online.

The trace of network activities includes IP address, online trait and privacy etc.

B. People's Information Privacy Right in E-commerce

People's information privacy right in E-commerce is a basic right in E-commerce era. It involves every part of the people's information collecting, transmitting, accessing and processing, such as:

(1) Right of privacy is not been peeped and intruded. It means that the privacy information shall not been peeped and intruded without being permitted.

(2) Right of privacy is not disturbed. It means that the network should be smooth and may not be intentionally interfered during online trading and online transfer

private information.

(3) Right of privacy is not been illegal collected, used and arbitrary spread. It means that the privacy information, such as personal basic data, special information, secret data, trust information, transaction information, trace of business activities online etc., is not been collected, used and spread without being permitted.

(4) Right of privacy is not been updated. It means that the privacy information, such as personal data, trust information and other important business information, should not be updated, tampered and distorted.

III. CAUSING REASONS ANALYSIS ON PEOPLE'S INFORMATION PRIVACY RIGHT IN E-COMMERCE

There four main reasons causing the problem of people's information privacy right in E-commerce. First is that the sense of security safeguard is weak. Second is that the E-commerce external environment is not perfect. Third is that the security measures of E-commerce system is not strong. Forth is that the malicious attacks come from various aspects [1].

A. The Sense on Security Safeguard in E-commerce is Weak.

In e-commerce, people is as the main part of business, but due to most of them have weak security awareness and the lack of the necessary security technical knowledge, and can not comply with or ignore safety operation, leading to private information disclosure and privacy be violated.

(1) Sense of security safeguard is weak. For examples, it lacks of the sense of security protection, adopts unsafe operations in E-commerce.

(2) There is no good security operation habit in E-commerce activity.

B. The External Environment of E-commerce is not Healthy.

The external environment of E-commerce includes laws and regulations, third-party services (such as bank, authentication center, credit management and credit service, logistic and distribution, etc.), human ethics, computer support technology and so on.

(1) On the aspect of laws and regulations related. Although the "Digital Signature Law" promulgated and implemented and the "Personal Information Protection Law" will published very soon in China, and there are many items in most of our laws and regulations. However, there are many shortages existed in the existing relevant laws and regulations, such as the laws and regulations is still not completely and does not form system, law enforcement strength and punishment intensity are not enough, lack of incentives, lag behind the development of E-commerce.

(2) On the aspect of third-party services. Due to the constraint of laws and regulations is not very strong and the standards and norms are imperfect, these will lead to make various phenomenon of invading people's privacy right, such as to obtain the account and password of the bank cards and credit cards of users by cheating, stealing, testing and many other means, and will lead to false identity authentication; credit information management and credit services are imperfect, and will lead to credit information has been tampered, credit level distortion.

(3) On the aspect of human ethics. Now, the social credit system is imperfect, the honest and trustworthy social climate is forming. So the fraud, theft, distortion, and wanton dissemination of people's privacy information occur sometimes.

(4) On the aspect of public network security. Some network corporations and Websites can not obey their promise to protect the customer's privacy while they provide network services, father more, they deliberately distribute, sell people's privacy information, and the result is that the people's information privacy right is harmed.

(5) On the aspect of privacy protection technology. Because the information on internet is transmitted by the router, so some illegal persons or organizations steal user privacy information by scanning the key nodes and tracking the transmitting activities. In addition, while data warehouse, data mining technology is rising and using quickly, some illegal persons or organizations process and utilize the personal information indiscriminately, and these make the people's privacy information revealed.

C. The E-commerce System has not adopted Effective Safeguard Measures.

Now, most of E-commerce systems have no any effective safeguard measure in program design, network and system architecture, data security connection and management, identification and authentication, firewall configuration, intrusion detection and prevention etc. This leads the people's privacy information disclosure, and the right of people's privacy has been infringed.

(1) There are many shortcomings on developing trading software and bad behaviors of application development. For examples, ASP, JSP, J+ +, J2EE, WebService, etc., all of them exit some shortcomings. This often leaves a lot of "backdoor" and provides the interface for illegal users to load the application module (such as Trojan horses, etc.).

(2) It has not taken physical isolation measures between intranet and extranet, network and storage media. This leads to an intruder invades the internal network through external network to get the privacy information stored in the internal computers. As well as, neglecting the hierarchical management of the security and the logical isolation of the internal network information system will lead the people's privacy information disclosed.

(3) It does not use the ODBC method to connect with databases, but use direct accessing will lead the database information disclosed.

(4) The accounts and passwords are too simple in user identification and authentication, the password, this is easy to be cracked or guessed.

(5) There is no firewall or the firewall setups very simple and unreasonable, this leads the illegal invaders to steal secrets easily bypass the firewall.

(6) The lack of proactive intrusion detection, identification, forensics and other security measures leads the variety of network intrusion rampant fraud invading the network and leads the privacy information disclosed.

D. The Malicious Attacks come from Various Aspects.

Now, the malicious attacks include worms, backdoors, Rootkits, DoS, and Sniffer, etc. In recent years, malicious attacks in E-commerce are becoming more intelligent, the attack tactics are escalating renovation, and harmful level is increasing. According to analyzing, the main malicious attacks on privacy information in E-commerce shows bellow [2].

(1) Monitor and password attacks. Because many of the agreements encryption or authentication technology have not adopted in every application-layer of E-commerce security system, and the user account and password information is transmitted in plain text format. This leads the attackers to make data monitor, to intercept confidential information transmitted by internet, public telephone network, line or installing receiving device in the range of electromagnetic radiation. Or, they infer useful information, such as bank account number, password and so on, by analyzing the volume of information flow, flowing direction, parameters of communication frequency.

(2) Network spoof attacks. By redirecting ARP cache communication data packets, rewrite the address mapping form the target machine's IP address to Mac, and leads the packet sent to the listener by the switch machine. This will lead privacy disclosed. The main methods of spoofing include Web spoofing attacks, TCP/IP spoofing, DNS spoofing, IP or name spoofing attacks.

(3) WWW attacks. This kind of attacks always use Java, ActiveX, JavaScript etc. to rewrite URL address and relative information to realize attacking.

(4) Trojan attacks. It is based on the network C/S principle, the attackers install C/S programs which communicate by ports on your computer, and the special programs will start when the computer operates in order to control the computer or steal important information.

(5) Buffer over flow attacks. Attackers can join attack codes, and enhance access right and control the computer when the buffer is overloaded if the buffer areas is overloaded and is not controlled.

(6) Rootkits attacks. Because the rookits is promised by attackers to access by backdoor, this will give a chance for attackers to start a Trojan to make attack or steal people's privacy information.

(7) Port attacks. The attackers can make attack by banding a Trojan to a legal port and get a legal ID, and enhance the right to get higher account and password.

(8) Sniffer backdoor attacks. The attackers can make attack by working under the hybrid / non-promiscuous mode.

(9) State manipulation attacks. The attackers can achieve the illegal visiting purpose by modifying the sensitive information, hidden form elements, and cookies in URL.

(10) SQL code embedded attacks. The attackers can make attack by inserting database query commands in the user input to realize database query, modify, and delete data and so on.

The harm of these malicious attacks is: attack the E-commerce system, to undermine the reliability of trading systems; disguise legal status, to undermine the authenticity of transaction identity; copy and theft of trade secrets, to undermine the confidentiality of information; tampers and delete information, to destruct the integrity of information; update and change information, to undermine the validity of information; invade the information system of certification department, to destruct non-repudiation of transactions.

Ⅳ. PROTECTION MEASURES OF PEOPLE'S INFORMATION PRIVACY RIGHT IN E-COMMERCE

How does prevent exposure the people's privacy information to protect their privacy right in E-commerce activities? Survey shows [3-4]: In the Internet users, there is 50% consider that increasing their own protection awareness is most important, 29.17% believe that installing a firewall and anti-virus software is important, 11.96% think that the important data is not online is a good idea, there are 6.88% consider that downloading the security patches regularly is very important. But the legal professions think that should speed up the construction of the relevant laws and regulations to protect the right of people's privacy in E-commerce. I believe that we should do the following four aspects well.

A. To Improve the Protection Awareness of People

(1) Should cultivate and improve the people's protection awareness

Because that the most privacy information disclosed events happen in "unconsciously", so it needs to strengthen the protection awareness and ability of privacy protection in E-commerce, such as education and training, to make peoples protect their own privacy right consciously.

(2) Should improve the people's security ability and make them have a good security operation habits.

The good security operation habits to protect privacy information in E-commerce includes: right use and setup Cookie, install privacy protection software and Cookie process software, protect password (it is not simple and general, do not access password), filter and mask threatening ActiveX, erase the traces of computer timely, hide or encrypt secret data, refuse the access come from threaten Website, install firewall, protect IP, use agent server, update the BUG of Windows generally, hide IP, plug the loopholes, erase "the documents saved" and "log files" and "attribute information of files" and "the history records in favorite".

B. *Improve the External Environment of E-commerce*

(1) Establish and improve the relevant laws and regulations to protect information privacy right of peoples

In China, the present laws setup on information privacy right is later than other country. The present laws are only relating to the Constitution, criminal law, civil commercial law and other, but there is no special law on protecting the right of people's information privacy. It should expect that the "Personal Information Protection Law" has entered the ranking of stage and will soon be promulgated and implemented.

In addition, the existing laws and regulations on privacy right protection are seriously lagging behind the rapid development of E-commerce. So we suggest that, while we setup laws, they should have a certain perspective, fully predict and estimate the development of E-commerce in future and the privacy violations that may occur; formulate relevant laws and regulations should take into account interests between information service provider and network application service provider, and peoples, and make be balance and coordination between them.

(2) Strengthen the integrity and moral setup, improve social credit system and credit management, to regulate credit information service

First is that it should strengthen education for peoples and make them respect for the privacy of others, and form a good social habit.

Second is that it should improve the social credit management system, solve the longstanding problems in the process of building a social credit system, and truly solve is credit information bottleneck.

Third is that it should improve credit information service management, evaluate the level and rate of credit service organizations, and make them have a high credit level first.

(3) Strengthen application study on network and information security technology applied in protecting privacy.

The technologies on protecting the people's information privacy right include digital certificate and encryption technology, P3P technology, firewall technology, invading checking technology, and information camouflage technology (such as digital watermarking, data hiding and data embedding) etc.

C. *Improve Security Ability of E-commerce System*

(1) E-commerce system developers should have good programming habits, take effective measures to seek to complete the procedure and software, eliminate the "back door" invasion of infringement behaviors (Trojan horses, etc.) to protect people's privacy information accessed without authorizing. In addition, they should update the system software and make "patchs" timely.

(2) Should adopt the physical isolating measures between internal network and external network, and between network and storages, to protect privacy information.

(3) Accessing databases should use ODBC dynamic link technology, but not use direct access method. This can reduce the choice of information in database disclosed.

(4) Should adopt long password in user's identity identification and authentication to increase the difficult of them cracked or guessed.

(5) Should install firewall software to prevent the illegal intruder to steal secrets.

(6) Should take a active intrusion detection, identification, forensics and other security defense measures to detect and screen a variety of network fraud intrusion.

D. *Prevent Malicious Attacks*

There are two steps to prevent malicious attacks. First is the malicious attacks detected. We can use detection technology to make real-time track, analyze and discover the malicious attacks in E-commerce. Second is adopting preventing measures to prevent malicious attacks.

(1) Detecting for the malicious attacks

The malicious attacks detection technologies include intrusion detection technology, trap technology and forensic technology [2].

Intrusion detection technology is based on Statistics and Fuzzy Logic to analyze. It is adapted to detect the malicious attacks that have anomalies. But it is difficult to detect network spoofs, Trojans and other hidden attacks.

Trap technology is based on making simulation network environment, setup loopholes to decoy attacks. It is adapted to the attacks which realize the gore by repeating test, such as sniffer backdoor. But it is difficult to detect embedding attacks and spoof attacks and denial service attacks.

Forensic technology is based on installing agents, creating diary log records to obtain proofs by analyzing. It is adapted to get proofs after attacked. But it is difficult to detect all kinds of present attacks.

(2) Prevent malicious attacks

We can adopt bellow technique measures to prevent the malicious attacks in E-commerce.

First is encryption technology. It is the main security preventing measure in E-commerce. It includes general encryption physic, data encryption standard (DES), symmetric encryption (private secret key cryptography system), asymmetric encryption (public key cryptography system), and symmetric secret key management.

Second is digital authentication and digital signature. We can use this technology to prevent data updated, confirm identification, prevent deny.

Third is identification center. It is used to provide identification identify and credit services.

Forth is to install firewall on the users' computers.

Fifth is SSL and SET security technology. These technologies are the key security technologies in payment system in E-commerce. We can use them to improve confidence, to guarantee information completion in trading, to improve security and reduce spoofs in E-commerce.

Sixth is security scan. It is an important technology in network security prevention. It is used in port scan, loophole scan to make intrusion detection.

Seventh is physical isolation. It is used to realize departing from internal and external, to avoid damaging from hacks and to protect privacy.

Eighth is to prevent monitoring and spoofing. We can use hard coding, IPSec, VPN and other encryption technologies to protect sensitive information.

Ninth is to prevent Web service attacks. It is need to strengthen the professional training for application developers to void various loopholes remained while they are developing software.

Tenth is XML technology. We can use XML technology to establish effective mechanisms to ensure the information flow safe transmitted, and to realize the trade security in E-commerce.

Eleventh is P3P technology. Using P3P technology in E-commerce can realize the safe transmitting of dynamic information and security protection of static privacy information, and prevent various attacks to steal the transaction information.

Twelfth is digital watermark technology. Because watermark has robustness, transparency, anti-aggressive properties, and low complexity, etc. We can use it to realize identifying and information hiding to protect information transmitted in E-commerce safe.

Thirteenth is voice or fingerprint identification technology. We can use both of voice/fingerprint and encryption technologies to realize double identification to improve the security of E-commerce, and it is effective for anti-repeat attacks.

Fourteenth is smart card and middleware technologies. We can use them to realize “hard program” to overcome the shortages of traditional security technologies.

Fifteenth is information camouflage. The camouflage can be divided into basic camouflage, watermark, data hiding and embedding technology [8]. We can use the special characters, such as hiding, security, correction and so on, to realize information hiding, right of information identification, identity identification etc.

In a word, in practice application, we should comprehensively use many of them to build integration security environment to improve the level of protecting privacy information safe in E-commerce, and to prevent malicious attacks come from every aspect.

Ⅴ. CONCLUTIONS

It is a complex system engineering to protect the people’s information privacy right in E-commerce. It not only involves the security problems of E-commerce privacy that is transmitted dynamically, but also includes the security problems of E-commerce privacy that is been storage statically, it is not merely a technical security problem but also includes the non-technical problems such as laws and regulations, strategy measures, credit system, commercial idea and so on. In the process of E-commerce developing, we want to solve the preventing problem of people’s privacy information right in E-commerce, to promote the health and fast developing of E-commerce, to protect the privacy right of both sides in E-commerce transaction, to promote the harmonious development of society. The first thing we must do is that to establish and perfect the corresponding laws and regulations and the social credit system, to build the honest and honor, health and harmonious humanities social environment. The second thing we must do is that to make the rules of safe operation for E-commerce, to draw up effective managing, strategies and measures of protecting privacy. The third thing we must do is that to improve the security performance of E-commerce software, to foster good habit of developing application programs, to strengthen the security measures of protecting the databases of E-commerce, to promote the studying on the technology protecting E-commerce privacy. Therefore, this article overcomes the one-sidedness existed in formerly studying such as "the heavy theory but light practice, the heavy qualitative but light quota, the heavy management but light technology", it has certain theoretical and practical significance for protecting the privacy right in E-commerce and for promoting health and fast developing of E-commerce.

Ⅵ. ACKNOWLEDGMENT

This work is supported by Philosophy and Social Sciences Foundation (07O05) of Guangdong Province, China.

REFERENCES

[1] Xiaoming Meng. “Study on safeguard measures of network privacy”. *Modern Library and Information Technology, Peking, China*. 2005(4), pp.92-95, 91.

[2] Xiaoming meng. “Study on detection and prevention technology for malicious attacks in E-commerce”. *Electronic Commerce Study, Peking, China*, 2006(4), pp.43-48.

[3] Licahng Guo. “How to protect privacy right using laws”. *Guangming Daily, Peking, China*, 22th. August, 2008.

[4] Tianwen Xu. “Principles and thought on privacy right in China”. *Journal of Zhuhai Municipal Administration Institute, Zhuhai, China*, 2007(2), pp.57-60.

[5] Yan Ceng, Bin Cheng, Zhonglin Liu. “Discuss network privacy right on the view of legal, thechnology and ethics”. *China Information Review, Peking, China*, 2006(2), pp.32-35.

[6] Wenjuan Wei. “Discuss on the legal protection of network privacy right”. *Legal System and Society*. Yunnan, China, 2007(1), pp.286.

[7] Xiao Hong. “Discuss on construction of network privacy right protection”. *Library Theory and Practice, Peking, China*, 2006(5), pp.38-39.

[8] Yixian Yang. “Information camouflage and security”. *Computer Security, Peking, China*, 2002(11), pp.50-53.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 488-491

On MAS-Based Automotive Electric Power Steering System Control Strategy and Architecture

Chuanyi Yuan [1], and Jingbo Zhao [2]
[1] School of Mechanical & Automobile Engineering, Jiangsu Teachers University of Technology
Changzhou 213001, China
Email: zhaojb1128@yahoo.com.cn
[2] School of Urban Railway Transportation, Soochow University, Suzhou 215021, China
Email: jbzhao@suda.edu.cn

***Abstract*—Electric Power Steering (Abbr. EPS) is a full electric system which reduces the amount of steering effort by directly applying the output from an electric motor to the steering system and has attracted much attention for their advantages. The constitutions and working principle of EPS system was introduced. Through the control requirements in EPS system, the MAS-based EPS control strategy and architecture was discussed and the basic function and behavior of every agent was represented. And according to the different working conditions, EPS system was divided into assist control model, return-to-center control model and damping control model, and corresponding flexible PID control algorithm, Fuzzy-PID control algorithm, Bang-Bang-PID control algorithm were designed. It has practical engineering significance to the design of EPS motor control strategy, to the improvement and optimization of EPS function and to the steering manipulation safety and provides an effective control method for EPS system.**

***Index Terms*—MAS; automotive; Electric Power Steering System; architecture; control strategy**

I. Introduction

Compared with the traditional hydraulic power steering device, Electric Power Steering has the advantages of safety, energy saving, environmental protection and so on. It has become the mainstream of power steering technology for passenger cars because of the important role of improvement of portability and the enhancement of handling stability and safety. EPS system is a complex non-linear dynamics containing different working conditions. The requirements of vehicle steering are high and involve many factors and it's difficult to adopt a single control strategy to coordinate the conflicts between the performance requirements under different conditions [1, 2]. In this paper, a multi-agent system method was introduced to deal with the control problem of EPS system. First, the architecture and working principle of EPS system was introduced. Then, through the control requirements in EPS system, the MAS-based EPS control strategy and architecture was discussed and the basic function and behavior of every agent was represented. And according to the different working conditions, EPS system was divided into assist control model, return-to-center control model and damping control model, and corresponding Flexible-PID control algorithm, Fuzzy-PID control algorithm, Bang-Bang-PID control algorithm were designed. It has practical engineering significance to the design of EPS motor control strategy, to the improvement and optimization of EPS function and to the steering manipulation safety and provides an effective control method for EPS system.

II. Working Principle of EPS System

A. Working principle of EPS system

Electric power steering system (Figure 1) is designed to use an electric motor to reduce effort by providing steering assist to the driver of a vehicle. It consists of a torque sensor, which senses the driver's movements of the steering wheel as well as the movement of the vehicle; an ECU, which performs calculations on assisting force based on signals from the torque sensor and vehicle sensor; a motor, which produces turning force according to output from the ECU; and a reduction gear, which increases the turning force from the motor and transfers it to the steering mechanism. By incorporating electronic stability control electric power steering systems can instantly vary torque assist levels to aid the driver in evasive manoeuvres and allow varying amounts of assistance to be applied depending on driving conditions.

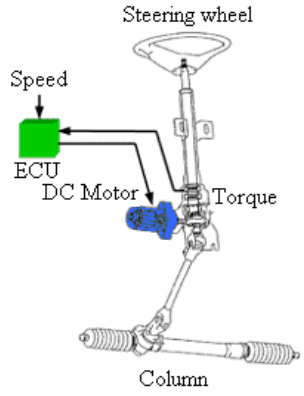


Figure 1. A column-type EPS system

EPS system has a slight advantage in fuel efficiency because it is no belt-driven hydraulic pump constantly

running, whether assistance is required or not, and this is a major reason for their introduction. Another major advantage is the elimination of a belt-driven engine accessory, and several high-pressure hydraulic hoses between the hydraulic pump, mounted on the engine, and the steering gear, mounted on the chassis which greatly simplifies manufacturing and maintenance [3].

B. EPS motor operation principle

The motor for EPS is a permanent magnetic field DC motor. Attached to the power steering gear assembly, it generates steering assisting force. Figure 2 illustrates the construction of a DC motor, consisting of a stator, a rotor, and a commutation mechanism [4, 5].

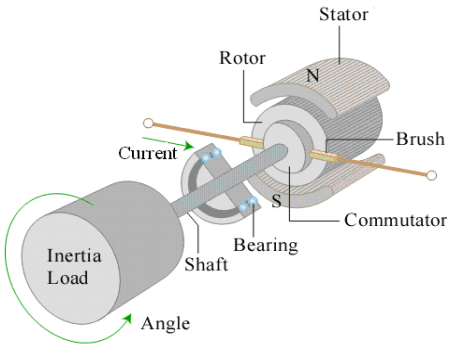


Figure 2. DC motor construction

The stator consists of permanent magnets, creating a magnetic field in the air gap between the rotor and the stator. The rotor has several windings arranged symmetrically around the motor shaft. An electric current applied to the motor is delivered to individual windings through the brush-commutation mechanism, as shown in the figure. As the rotor rotates the polarity of the current flowing to the individual windings is altered. This allows the rotor to rotate continually. Figure 3 is the schematic of the electric circuit, including the windings resistance R and inductance L.

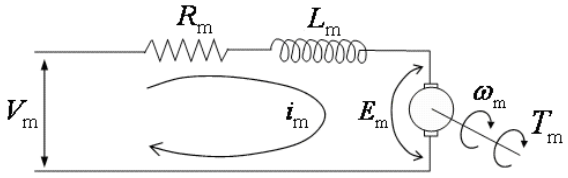


Figure 3. Electric circuit

We can get the transfer function of a DC motor, and the mathematical model is given by

$$\begin{cases} V_m = R_m i_m + L_m \dfrac{\mathrm{d}i_m}{\mathrm{d}t} + e_m \\ T_m = K_t i_m \\ T_m = T_L + T_f \\ e_m = K_e \omega_m \end{cases} \quad (1)$$

The mathematical model in the frequency domain are

$$\begin{cases} V_m(s) = R_m I_m(s) + L_m s I_m(s) + E_m(s) \\ J_m s\omega_m(s) = -B_m \omega_m(s) - T_L(s) + T_m(s) \\ E_m(s) = K_m \omega_m(s) \\ T_m(s) = K_m I_m(s) \end{cases} \quad (2)$$

If set $T_L = 0$ and relate V_m to T_m, there is

$$T_m^{(1)}(s) = \frac{K_m(J_m s + B_m)}{(L_m s + R_m)(J_m s + B_m) + K_m^2} V_m(s) \quad (3)$$

If set $V_m = 0$ and relate T_L to T_m, there is

$$T_m^{(2)}(s) = \frac{K_m^2}{(L_m s + R_m)(J_m s + B_m) + K_m^2} T_L(s) \quad (4)$$

There is the matrix transfer function as following

$$\begin{aligned} T_m(s) &= T_m^{(1)}(s) + T_m^{(2)}(s) \\ &= \frac{K_m(J_m s + B_m)}{(L_m s + R_m)(J_m s + B_m) + K_m^2} V_m(s) \\ &+ \frac{K_m^2}{(L_m s + R_m)(J_m s + B_m) + K_m^2} T_L(s) \quad (5) \\ &= \begin{bmatrix} \dfrac{K_m(J_m s + B_m)}{(L_m s + R_m)(J_m s + B_m) + K_m^2} \\ \dfrac{K_m^2}{(L_m s + R_m)(J_m s + B_m) + K_m^2} \end{bmatrix}^{-1} \begin{bmatrix} V_m(s) \\ T_L(s) \end{bmatrix} \end{aligned}$$

III. WORKING PRINCIPLE OF EPS SYSTEM

A. MAS-based EPS control structure

As is known to all, EPS system is a complex non-linear dynamics containing different working conditions. The requirements of vehicle steering are high and involve many factors and it's difficult to adopt a single control strategy to coordinate the conflicts between the performance requirements under different conditions. An agent is a computer system that is capable of independent action on behalf of its user or owner, that is, which figures out what needs to be done to satisfy design objectives, rather than constantly being told. And a multi-agent system is one that consists of a number of agents, which interact with one-another. In the most general case, agents will be acting on behalf of users with different goals and motivations. To successfully interact, they will require the ability to cooperate, coordinate, and negotiate with each other, much as people do. A multi-agent system method should be introduced to deal with the control problem of EPS system.

MAS-based EPS control structure consists of Data Acquisition Agent, Condition Monitoring Agent, Controller Selection Agent, Flexible-PID Control Agent,

Fuzzy-PID Control Agent, Bang-Bang-PID Control Agent, PWM Drive Agent and Stability Control Agent which is shown in Figure 4.

Data Acquisition Agent is used to obtain the EPS system's data and stored in the database for Condition Monitoring Agent and Controller Selection Agent calls for determining the choice of controller. Stability Control Agent is used to monitor and ensure that all controllers and the whole EPS system's stability. PWM Agent is used to adjust the PWM duty cycle to obtain the magnitude and direction of the motor [6, 7].

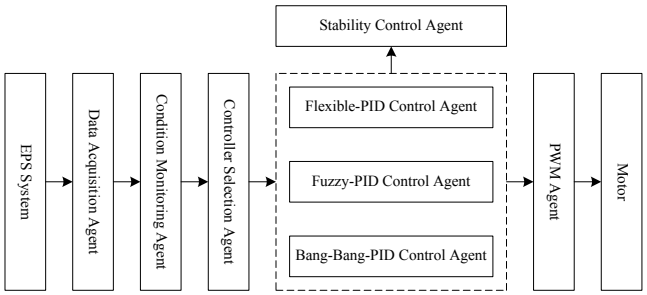


Figure 4. MAS-based EPS control structure

B. EPS control structure

The block diagram of control strategy of EPS system is shown in Figure 5. The assist characteristic unit determines the reference current to the motor based on the driving conditions, and the controller computes the control signal which minimizes the error between and the actual current. The EPS controller conducts a search for data according to a table lookup method based on the signals input from each sensor and carries out a prescribed calculation using this data to obtain the assist force.

Another important technology is the generation of Pulse width modulation (PWM) signal which is employed in a wide variety of applications, ranging from measurement and communications to power control and conversion [4, 5].

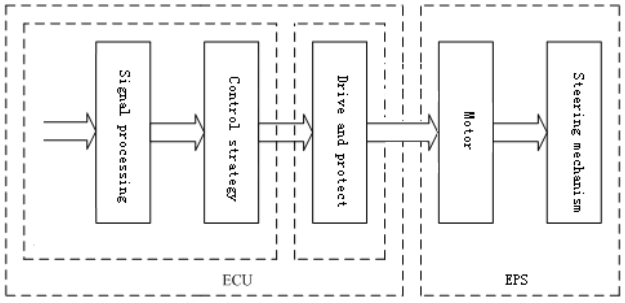


Figure 5. EPS control strategy

In the assist control model and for Flexible-PID Control Agent, the target motor current which is proportional to the motor assist torque is determined from the signal output from the torque sensor, and the Flexible-PID controller is performed so that there is no difference between this target current value and the value detected through feedback from the current sensor (Figure 6).

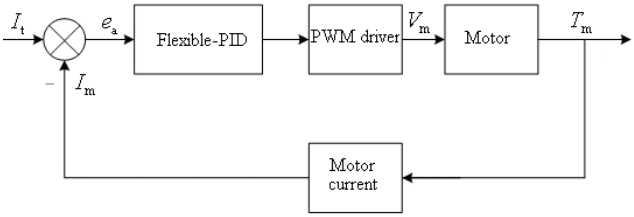


Figure 6. Flexible-PID controller

In the return-to-center control model and for the Fuzzy-PID Control Agent, the vehicular control requirements are when at low speed the return curve must pass back to the starting point, and when at high speed the allowed residual angle was not allowed to exceed 5° (Figure 7).

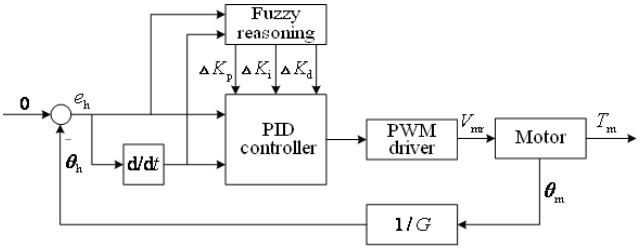


Figure 7. Fuzzy-PID controller

In the damping control model and for the Bang-Bang-PID Control Agent, the control structure was adjusted between the Bang-Bang controller and PID controller so as to make the error reduced by using the shortest optimal control problem (Figure 8).

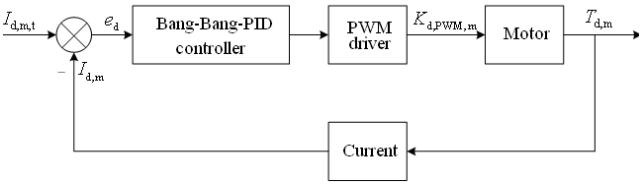


Figure 8. Bang-Bang-PID controller

Ⅳ. CONCLUSIONS

For EPS system, assist motor, torque sensor and Electronic Control Unit are the three key components. The core of the controller is control strategy which not only reflects the EPS control functional requirements, the level of adaptive capacity and intelligence as well as the key bottleneck and technology of independent research and development.

The MAS-based EPS control strategy shows the advanced nature, superiority and feasibility of the theoretical analysis and engineering application of EPS control system. In different control model, the switching and function of each controller were effective and could meet the real-time control demand under different working conditions. It has practical engineering significance to the design of EPS motor control strategy, to the improvement and optimization of EPS function and

to the steering manipulation safety and provides an effective control method for EPS system.

REFERENCES

[1] Jiang Haobin et al.Hardware design and experiment research of automotive electric power steering system. The 3rd China-Japan Conference on Mechatronics 2006 Fuzhou,2006,68-71.

[2] Aly Badawy et al.Modeling and analysis of an electric power steering system.SAE paper 1999-01- 0399.

[3] Zhao Jingbo, Chen Long, Jiang Haobin, et al. Design and full-car tests of electric power steering system. Computer and Computing Technologies in Agriculture. United States: SPRINGER, 2008: 729-736.

[4] Ronald K. Jurgen. Automotive electronics handbook [M], Second edition, McGraw-Hill, Inc, 1999.

[5] Zhao Jingbo. Research on Automotive EPS Hybrid Control System and its Theory, Design and Realization. Ph.D. Dissertation, Jiangsu University, 2009.

[6] B. P. Zeigler, T. G. Kim, and H. Praehofer. Theory of Modeling and Simulation.Academic Press, Inc., Orlando, FL, USA, 2000.

[7] R. E.Wray and R. M. Jones. An introduction to soar as an agent architecture. In R. Sun,editor, Cognition and Multi-Agent Interaction: From Cognitive Modelling to Social Simulation, pages 53 – 78. Cambridge University Press, 2005.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 492-495

An Approach of time-delay Switch Control for CSC Inventory System

Chunling Liu , Cheng Chen , and Apin Yuan
School of electronics and information, Wuhan University of Science & Engineering
E-mail:chunringliu5@yahoo.com.cn

Abstract— **The application of switch contol methodology in complex system such as industry of mechinery, electronics is widespread. In this paper the robust control system with time-delay based on switch theory (STDRCS) is presented to manage the inventory of cluster supply chains（CSC）. The hybrid control approach might be effectively used to overcome model uncertainties and external disturbance. And the robust inventory control strategy is aimed to find the optimal decision variables to weaken the bullwhip effect in cluster supply chains. It is proved that this method is more effective and efficient to tune the controller in face of changing environment.**

Index Terms— **inventory management; robust inventory; online switched system; cluster supply chains**

I. INTRODUCTION

Improved operation of supply chains for manufacturedgoods is worth billions of dollars to the national economy; effective inventorymanagement plays an important role in this regard.The use of optimization techniques in the management ofsupply/demand networks began with the development of the classical economic order quantity approach. Later developments include approaches for determining optimalbase stock levels in "order-up-to" policies.Cluster supply chain (CSC), different from traditional single-chain supply chain, is located in the industrial cluster region, with the relation of "supply-client", through the link of formal or informal contract of 'trust and commitment', formed by organizations containing different firms of the same industry such as research organizations, supplier, manufacturer, wholesaler, retailer, and even end users. Cluster supply chain system is made of a couple of paralleled single supply chains in the agglomeration location, not only do all enterprises in one single supply chain cooperate one another internally, but cooperation and coordination exist across different single supply chains externally as well[1].

Nowadays, numerous literatures have been devoted to study of inventory control models in supply chain, and made fruitful progress. However, few of these involve across-chain inventory management which objectively exists among cluster supply chain in many industrial cluster locations[4]. In addition, dynamic uncertain environments and integrated control for complex systems require developing a model of combining inventory management and manufacturing process[2][3]. So, the switched robust control under time-delay (STDRCS), one of hybrid controls, is introduced to this paper for establishing the model, therefore, we focus on across-chain inventory in cluster supply chains to analyze its bullwhip effect and approach of implementing it. The following parts are arranged in such way: section II briefly describes switched robust control under time-delay. The section III explore the hybrid optimization and decision approaches. At last, an example is used to show the solving procedure in section IV.

II. SWITCHED ROBUST CONTROL WITH TIME-DELAY

Switch Robust Control with time-delay (STDRCS) is a robust control method controlling system structure uncertainty. The switched controllers considered in this paper are based on the state-space model below:

$$\hat{x}(k+1) = A_i\hat{x}(k) + \tilde{A}_i\hat{x}(k-h) + B_2\hat{u}(k) + B_1\omega(k) \quad (1)$$

Where, $i \in I, I = \{1,2,\cdots,N\}$ is switched rules, $A_i \in R^n$, $\tilde{A}_i \in R^n$, h is the unknown time-delay integer constant, $k \geq h$, $\hat{x}(k-h)$ express unreached goods ordered before the k period. $\hat{x}(\bullet) \in R^n$ is state vector, $\hat{u}(\bullet) \in R^n$ is control vector, $\omega(\bullet) \in R^n$ is external disturbance,

Given the general definition of quadratic cost function

$$J = \sum_{k=0}^{\infty}[\hat{x}^T(k)Q\hat{x}(k) + \hat{u}^T(k)R\hat{u}(k)] \quad (2)$$

Where, $0 < Q = Q^T \in R^{n\times n}$ and $0 \leq R = R^T \in R^{n\times n}$ are respectively state and control weight matrix.

Define the subsidiary output sign as the following

$$z(k) = C\hat{x}(k) + D\hat{u}(k) \quad (3)$$

Where, $C = \begin{bmatrix} Q^{1/2} & 0 \end{bmatrix}^T$, $D = \begin{bmatrix} 0 & R^{1/2} \end{bmatrix}^T$, then the quadratic index（2）can be represented as

$$J = \sum_{k=0}^{\infty} z^T(k)z(k) = \|z\|_2^2 \quad (4)$$

For the uncertain discrete system (1) and (3), introducing the state feedback control as

$$\hat{u}(k) = K_{1i}\hat{x}(k) + K_{2i}\hat{x}(k-h) \quad (5)$$

Accordingly, the closed-loop system may be represented:

$$\begin{cases} \hat{x}(k+1) = \sum_{m=1}^{s}\theta_{im}(A_i + B_2K_{1i})\hat{x}(k) + (\tilde{A}_i + B_2K_{2i})\hat{x}(k-h) + B_1\omega(k) \\ \hat{z}(k) = (C + DK_{1i})\hat{x}(k) + DK_{2i}\hat{x}(k-h) \end{cases} \quad (6)$$

In above, the SDRC controller selects the input $u(k)$ by solving the following optimization problem

AP-PROC-CS-09CN005

$$J = min\left(\sum_{k=0}^{\infty} z^T(k)z(k) - \gamma^2 \|\omega\|_2^2 \right) \quad (7)$$

Definition: for time-delay switched system (6), if the selected Lyapunov function $V(\hat{x}(k))$ along the difference of system (6) satisfy $\Delta V(\hat{x}(k)) < 0$, then the system is robust quadraticly stable

Theorem 1 for time-delay switched system (5), and given constant γ, If there existing positive definite matrix W_i and V, existing matrix G_{1i} and G_{2i}, thus $\forall(i,j) \in I = \{1,2,\cdots,N\}$

$$\begin{bmatrix} -V & W_i & 0 & 0 & 0 & 0 \\ W_i & -W_i & 0 & 0 & (A_iW_i + B_2G_{1i})^T & (CW_i + DG_{1i})^T \\ 0 & 0 & -2W_i + V & 0 & (\tilde{A}_iW_i + B_2G_{2i})^T & (DG_{2i})^T \\ 0 & 0 & 0 & -\gamma^2 I & B_1^T & 0 \\ 0 & (A_iW_i + B_2G_{1i}) & (\tilde{A}_iW_i + B_2G_{2i}) & B_1 & -W_j & 0 \\ 0 & (CW_i + DG_{1i}) & DG_{2i} & 0 & 0 & -I \end{bmatrix} < 0 \quad (8)$$

then the system existing state-setback control law with memory (formula (5)), $K_{1i} = G_{1i}W_i^{-1}, K_{2i} = G_{2i}W_i^{-1}$ and for arbitrary switched signal, the closed-loop system has the H∞performance γ.

Proof: for subsystem i, defy Lyapunov function as

$$V(k) = \hat{x}^T(k)P_i\hat{x}(k) + \sum_{\tau=1}^{h} \hat{x}^T(k-\tau)\Gamma\hat{x}(k-\tau)$$

where, P_i, Γ is positive definite matrix, then $V(k)$ is positive definite function。By definition and formula（5）, the sufficient condition that the closed-loop system（6）is robust quadratic stable and having the H∞performance γ. Thus

$$\begin{aligned} &V(k+1) - V(k) + \hat{z}^T(k)\hat{z}(k) - \gamma^2\omega^T(k)\omega(k) \\ &= \hat{x}^T(k+1)P_j\hat{x}(k+1) - \hat{x}^T(k)P_i\hat{x}(k) + \hat{x}^T(k)\Gamma\hat{x}(k) - \hat{x}^T(k-h)\Gamma\hat{x}(k-h) - \gamma^2\omega^T(k)\omega(k) \\ &= \begin{bmatrix} \hat{x}(k) \\ \hat{x}(k-h) \\ \omega(k) \end{bmatrix}^T \left\{ \begin{bmatrix} -P+\Gamma_i & 0 & 0 \\ 0 & -\Gamma & 0 \\ 0 & 0 & -\gamma^2 I \end{bmatrix} + \begin{bmatrix} (C_i + DK_{1i})^T \\ (DK_{2i})^T \\ 0 \end{bmatrix} \begin{bmatrix} C_i + DK_{1i} & DK_{2i} & 0 \end{bmatrix} \right. \\ &\left. + \begin{bmatrix} (A_i + B_2K_{1i})^T \\ (\tilde{A}_i + B_2K_{2i})^T \\ B_1^T \end{bmatrix} P_j \begin{bmatrix} (A_i + B_2K_{1i})^T & (\tilde{A}_i + B_2K_{2i})^T & B_1 \end{bmatrix} \right\} \begin{bmatrix} \hat{x}(k) \\ \hat{x}(k-h) \\ \omega(k) \end{bmatrix} < 0 \end{aligned}$$

By the Shur performance of matrix，existing

$$\begin{bmatrix} -P_i + \Gamma & 0 & 0 & (A_i + B_2K_{1i})^T & (C + DK_{1i})^T \\ 0 & -\Gamma & 0 & (\tilde{A}_i + B_2K_{2i})^T & (DK_{2i})^T \\ 0 & 0 & -\gamma^2 I & B_1^T & 0 \\ (A_i + B_2K_{1i}) & (\tilde{A}_i + B_2K_{2i}) & B_1 & -W_j^{-1} & 0 \\ C + DK_{1i} & DK_{2i} & 0 & 0 & -I \end{bmatrix} < 0 \quad (9)$$

multiply the above formula from left and right separately with matrix $diag([P_i^{-1}, P_i^{-1}, I, I, I])$, and let $W_i = P_i^{-1}, W_j = P_j^{-1}, V = \Gamma^{-1}$, then existing

$$\begin{bmatrix} -W_i + W_iV^{-1}W_i & 0 & 0 & W_i(A_i + B_2K_{1i})^T & W_i(C + DK_{1i})^T \\ 0 & -W_iV^{-1}W_i & 0 & W_i(\tilde{A}_i + B_2K_{2i})^T & W_i(DK_{2i})^T \\ 0 & 0 & -\gamma^2 I & B_1^T & 0 \\ W_i(A_i + B_2K_{1i}) & W_i(\tilde{A}_i + B_2K_{2i}) & B_i & -W_j & 0 \\ CW_i + DK_{1i}W_i & DK_{2i}W_i & 0 & 0 & -I \end{bmatrix} < 0 \quad (10)$$

For $V > 0$, there is $(V - W_i)^T V_i^{-1}(V - W_i) \geq 0$, that is $V - W_i - W_i + W_iV_i^{-1}W_i \geq 0$, therefore there is $+W_iV_i^{-1}W_i \geq 2W_i - V_i$。So there exists

$$\begin{bmatrix} -W_i + W_iV^{-1}W_i & 0 & 0 & W_i(A_i + B_2K_{1i})^T & W_i(C + DK_{1i})^T \\ 0 & -W_iV^{-1}W_i & 0 & W_i(\tilde{A}_i + B_2K_{2i})^T & W_i(DK_{2i})^T \\ 0 & 0 & -\gamma^2 I & B_1^T & 0 \\ A_iW_i + B_2K_{1i}W_i & \tilde{A}_iW_i + B_2K_{2i}W_i & B_1 & -W_j & 0 \\ CW_i + DK_{1i}W_i & DK_{2i}W_i & 0 & 0 & -I \end{bmatrix}$$
$$\leq \begin{bmatrix} -W_i + W_iV^{-1}W_i & 0 & 0 & W_i(A_i + B_2K_{1i})^T & W_i(C + DK_{1i})^T \\ 0 & -2W_i + V_{ii} & 0 & W_i(\tilde{A}_i + B_2K_{2i})^T & W_i(DK_{2i})^T \\ 0 & 0 & -\gamma^2 I & B_1^T & 0 \\ A_iW_i + B_2K_{1i}W_i & \tilde{A}_iW_i + B_2K_{2i}W_i & B_1 & -W_j & 0 \\ CW_i + DK_{1i}W_i & DK_{2i}W_i & 0 & 0 & -I \end{bmatrix}$$

In the above formulation, if the right matrix is passive definite, then the left one must be passive definite. So for arbitrary switched signal, the right matrix is a sufficient condition of robust quadratic stability and meeting H∞ performance γ for system (6). By Schur performance of matrix, the following formula can be made.

$$\begin{bmatrix} -V & W_i & 0 & 0 & 0 & 0 \\ W_i & -W_i & 0 & 0 & W_i(A_i + B_2K_{1i})^T & W_i(C + DK_{1i})^T \\ 0 & 0 & -2W_i + V & 0 & W_i(\tilde{A}_i + B_2K_{2i})^T & W_i(DK_{2i})^T \\ 0 & 0 & 0 & -\gamma^2 I & B_1^T & 0 \\ 0 & (A_iW_i + B_2K_{1i}W) & (\tilde{A}_iW_i + B_2K_{2i}W_i) & B_1 & -W_j & 0 \\ 0 & (CW_i + DK_{1i}W_i) & DK_{2i}W_i & 0 & 0 & -I \end{bmatrix} < 0$$

Let $G_{1i} = K_{1i}W_i, G_{2i} = K_{2i}W_i$, then formula（8） can be obtained. Then the theorem is proved.

III. INVENTORY CONTROL IN CLUSTER SUPPLY CHAINS

The inventory system of cluster supply chain in this section is composed of two single-chain supply chains which encompass one manufacturer and one retailer (showed in fig.1) and manufacture character-equal substitutable product.

Suppose x_1, x_2 represent inventory level of retailer and manufacturer in SC1 respectively, x_3, x_4 represent inventory level of retailer and manufacturer in SC2 respectively. Suppose u_1, u_2 represent order of retailer and manufacturer in SC1 respectively, whereas, u_3, u_4 represent order in SC2 respectively. ξ_1, ξ_2 represent the market demand of SC1 and SC2.

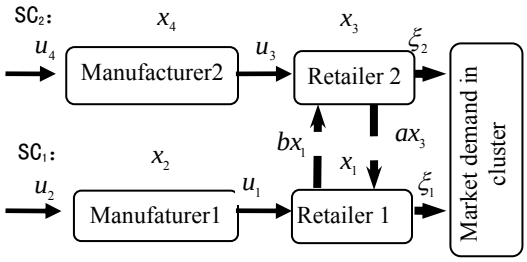


Fig 1 structure of across-chain inventory cooperation in CSC

In practice, the two supply chains maintain long-term cooperation so as to enlarge the whole demand. When the demand from customers of retailer 1 increases sharply and suddenly, then the retailer 2 may transship inventory to retailer 1 for its emergent need, the supply quantity as $a\hat{x}_3(0<a\le 1)$; when the uncertain demand from customer of retailer 2 increases sharply and suddenly, vise visa, the supply quantity as $b\hat{x}_1(0<b\le 1)$.

Thus, regarding the inventory state as the state variable, the inventory model may be defined as[4]

$$\boldsymbol{x}(k+1)=\boldsymbol{A}_i\boldsymbol{x}(k)+\boldsymbol{B}_1\boldsymbol{\xi}(k)+\boldsymbol{B}_2\boldsymbol{u}(k) \qquad (11)$$

Where, $A_i=\begin{bmatrix}1-b & 0 & a & 0\\ 0 & 1 & 0 & 0\\ b & 0 & 1-a & 0\\ 0 & 0 & 0 & 1\end{bmatrix}$, $B_1=\begin{bmatrix}-1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & -1 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}$,

$B_2=\begin{bmatrix}1 & 0 & 0 & 0\\ -1 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & -1 & 1\end{bmatrix}$, $\xi(k)=\begin{bmatrix}0\\ \xi_1(k)\\ 0\\ \xi_2(k)\end{bmatrix}$

The demand from the customer is divided into two parts of being certain and uncertain ones:

$$\xi_i(k)=d_i(k)+\omega_i(k) \qquad (12)$$

Thus, in seek of demand disturbance, the inventory variable (state variable) and order (control variable) are all disturbed in the cluster supply chain system (11).

Suppose the standard values of inventory vector, order are $\boldsymbol{x}_s,\boldsymbol{u}_s$, then the error system of cluster supply chain is

$$\hat{\boldsymbol{x}}(k+1)=\boldsymbol{A}_i\hat{x}(k)+\boldsymbol{B}_1\boldsymbol{\omega}(k)+\boldsymbol{B}_2\hat{\boldsymbol{u}}(k) \qquad (13)$$

The switched rules are set as

$$S=\begin{cases}a=0,0<b\le 1 \ \ when\ \omega_{1,k}>2S_{11} \ \text{and}\ \omega_{2,k}<2S_{11} \\ \qquad\qquad \text{and}\ \hat{x}_{21,k}>S_{21}\\ b=0,0<a\le 1 \ \ when\ \omega_{2,k}>2S_{21} \ \text{and}\ \omega_{1,k}<2S_{11} \\ \qquad\qquad \text{and}\ \hat{x}_{11,k}>S_{11}\\ a=b=0,\text{others}\end{cases} \qquad (14)$$

Where, S_{11},S_{21} are respectively secure inventory level of retailers in SC1 and SC2. The system satisfies $a\bullet b=0$ at any time, that is to say, the transshipment between the retailers may not happen, but they cannot mutually replenish at the same time.

The bullwhip effect is described as the proportion of the sum of inventory and order fluctuation to the terminal demand fluctuation, the definition may be showed in [4] namely

$$r_k=[(\hat{\boldsymbol{x}}_k)^T\boldsymbol{Q}\hat{\boldsymbol{x}}_k+(\hat{\boldsymbol{u}}_k)^T\boldsymbol{R}\hat{\boldsymbol{u}}_k]/(\boldsymbol{\omega}_k)^T\boldsymbol{S}\boldsymbol{\omega}_k \qquad (15)$$

Where Q,R,S are set symmetrical positively defini te weighted matrix. The parameter r_k describes bullw hip effect in cluster supply chains. The bullwhip effe ct becomes stronger with increase in r_k, while weake r with decrease in r_k.

For some external disturbance $\omega(k)$, if the controlled output $z(k)$ always maintains small level in the system, then the system with such index present "better" performance. In this case, the controlled output is less influenced by both external disturbance, and the capacity of restraining disturbance in the system appears stronger.

Thereby, the solver satisfying performance index (EQ. 8) surely guarantees minimizing bullwhip effect. The solving process of robust controller may be obtained with given model in section 2. However, how to switch among multiple modes must depend on one online decision-making system which will be introduced in section 4.

IV. SIMULATION TEST CASE

Two-echelon cluster supply chains are taken as instance in fig.1 and the switched robust inventory model (EQ. 1 and EQ.3) is given in the former section.

Suppose the standard system of cluster supply chains are

$x_{1s}=[1.2,1.3,1.5,1.55]'\text{Kton}$,

$u_{1s}=[1.2,1.35,1.6,1.7]'\text{Kton}$

And suppose the initial values of order error are zero, but the initial values of stock error are

$\hat{x}_0=[-0.1,0.1,0.35,0.1]'\text{Kton}$

The demand disturbances in the two supply chains are random. Suppose the system face demand disturbance shown in Fig.3. The retailer in SC1 has larger need at period of k=0 and k=15 (namely, $\omega_1>2S_{11}$). The other supply chain will provide emergent inventory supply for the retailer in SC1 by contract when meeting the condition of inventory transshipment. Suppose the supply ratio is a=0.8. The supply chain 2 faces larger demand fluctuation at k=14 in Fig 3, but the across-chain transshipment condition is not met (namely, $\omega_1<2S_{11}$), so supply chain 2 replenishes inventory only by single channel.

The subsystem 1 is set as system without across-chain inventory cooperation, the according parameters are

$A_1=\begin{bmatrix}0.8 & 0 & 0 & 0\\ 0 & 0.8 & 0 & 0\\ 0 & 0 & 0.8 & 0\\ 0 & 0 & 0 & 0.8\end{bmatrix}$, $\tilde{A}_{11}=\begin{bmatrix}0.2 & 0 & 0 & 0\\ 0 & 0.2 & 0 & 0\\ 0 & 0 & 0.2 & 0\\ 0 & 0 & 0 & 0.2\end{bmatrix}$,

The subsystem 2 is set as system with across-chain inventory cooperation (the retailer of SC2 supplies inventory for retailer of SC1), the according parameters are)

$$A_2 = \begin{bmatrix} 0.8 & 0 & 0.8 & 0 \\ 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.8 \end{bmatrix}, \quad B_{21} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix},$$

$\tilde{A}_1 = \tilde{A}_1, \tilde{A}_2 = \tilde{A}_2$.

$B_1 = diag([-1,0,-1,0])$, suppose $Q = R = S = I_4$.

The system is switched to mode 2 (subsystem 2 is determined) in periods of k=1、k=16, while mode 1 in other periods by results of online decision-making module. Set $\gamma = 3$, according to the condition that the system has certain performance of restraining disturbance and must satisfy robust stability (EQ. 8, 9).

In supply chain, the inventory fluctuation and order fluctuation at every tier can only maintain stable within a small scope in seek of demand disturbance. The simulation and computing are carried out by emulating online decision in this section, given the real demand disturbance shown in Fig. 3. In order to analyze the different varying trend of related parameters between considering time-delay and without considering time=delay, the system made simulation and computing separately in the two situations (Fig.4).

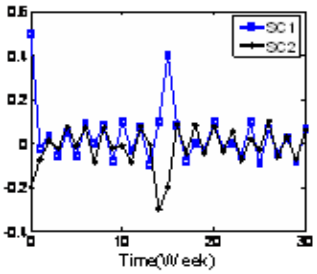
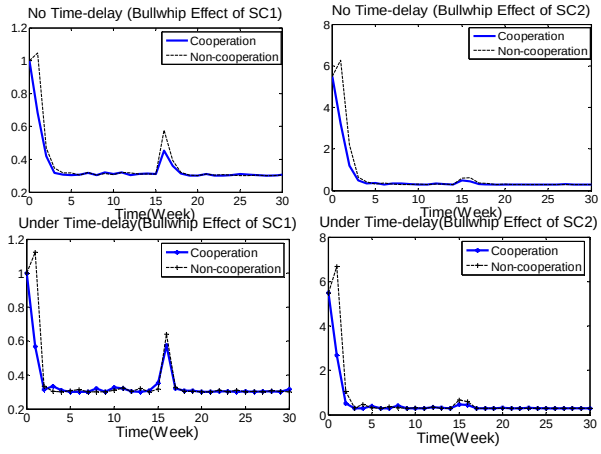


Fig. 3 Demand disturbance in CSC



(a)Supply chain 1 (b)Supply chain 2

Fig. 4 Variation tendency of bullwhip effect

It is shown from Fig.4 that the system may tend to be stable within a small scope by exerting H_∞ control with time-delay considered, while the effect is better when existing across-chain inventory cooperation. One evident reason for this result is that the order fluctuation is largely dwindled by across-chain inventory cooperation, thus the bullwhip effect can be better weakened.

Additionally, it is proved by simulation in this section that the online decision-making system of cluster supply chain with time-delay consideration can optimize the whole system in multiple periods, and then take on quicker response to complex market.

It is also shown from the figures that the varying trend of some variables in the system with time-delay consideration is inferior slightly to that of without time-delay because the system decision was conservative for time-delay. On the other hand, the transship quantity across chains is less than that of none time-delay, because the decision is based on the current inventory fluctuation and without stock in transit considered. However, the phenomenon of time-delay is objectively existing and the system with time-delay consideration conforms to reality. Thus it is necessary that the across-chain transshipment problem of stock with time-delay consideration is discussed in multiple views.

ACKNOWLEDGMENT

The paper is supported by China Educational Ministry Program (08JA630063), NASFC (70502029), China Post-doctoral Special Program (200801312)), The Educational Department Program of Hubei Province (Z20081702, B200717001) and National CIMS/863 (2009AA04Z153).

REFERENCES

[1] J. Z. Li. Inventory Management in Cluster Supply Chain. Press of Chinese Economy, 2006

[2] E. Lopez, B. Ydstie, I. E. Grossmann. A model predictive control strategy for supply chain optimization. Computers & Chemical Engineering, 27(2003), PP.1201-1218

[3] H. S. Sarjoughian, D. P. Huang. Hybrid Discrete Event Simulation With Model Predictive Control For Semiconductor Supply-chain Manufacturing. Proceedings of the 2005 Winter Simulation Conference, PP.256-265

[4] C. L. Liu, J. Z. Li. Research on H_∞ Control of Bullwhip Effect in Cluster Supply Chains Based on Cooperation between Two single Chains. Chinese Management Science, 2007, 15(1), PP. 41-46

[5] Daafouz, P. Riedinger, and C. Lung. Stability analysis and control synthesis for switched systems: a switched Lyapunov function approach. IEEE Trans. Automat. Control, 2002, 47(11):1883-1887

[6] S. Muller. http://www.held-mueller.de/JMatLink/

[7] Y. Zhang, P. Sen. & G. Hearn. An on-line trained adaptive neural controller. *IEEE Control Systems Magazine*, 1995, 15(5)

[8] R. Brown. Smoothing, forecasting, and prediction of discrete time series, Prentice-Hall, 1962

[9] S. Axsäter. Modelling emergency lateral transshipments in inventory system. *Management Science*, 1990, 36(11):1329-1338

[10] G.Tagaras, M Cohen. A Pooling in two-location inventory systems with non-negligible replenishment lead times. *Management Science*, 1992, 38(8):1067-108

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 496-499

From Graphical Model in UML Activity Diagrams to Formal Specification in Event B for Workflow Applications Modeling

Ahlem Ben Younes[1] , and Leila Jemni Ben Ayed[2]
[1] Research Unit of Technologies of Information and Communication (UTIC)- ESSTT-Tunisia
Ahlem.benyounes@fst.rnu.tn
[2] Research Unit of Technologies of Information and Communication (UTIC)- ESSTT-Tunisia
Faculty of Science of Tunis
Leila.jemni@fsegt.rnu.tn

***Abstract* –The lack of a precise semantics for UML AD makes the reasoning on models constructed using such diagrams infeasible. However, such diagrams are widely used in domains that require a certain degree of confidence. Due to economical interests, the business domain is one of these. To enhance confidence level of UML AD, this paper provides a formal definition of their syntax and semantics. The main interest of our approach is that we chose UML AD, which are recognized to be more tractable by engineers. We outline the translation of UML AD into Event B in order to verify functional properties of workflow models (such as deadlock-inexistence, liveness, fairness) automatically, using the B powerful support tools like B4free. We propose a solution to specify time in Event B, and by an example of workflow application, we illustrate the proposed technique.**

***Index Terms*—Specification, Formal verification, Validation, UML, Event B, workflow application**

I. INTRODUCTION

The work presented in this paper is part of our works [2] [13] that aims at providing specification and verification technique for workflow applications. A workflow is an operational business/scientific process. In order to represent workflows in an intuitive and practical way with an standard language, we have chosen UML AD. The modeling process is not addressed in this paper; more details about it can be found in [2]. However, the fact that UML lacks a precise semantics is a serious drawback of UML-based techniques. Also, UML AD is not adapted to the verification of workflow applications. In this paper, our goal is to provide a specification and verification technique for workflow applications using UML AD endowed with timed characteristics and synchronisation aspects. In fact, in the business domain, the emphasis is made on the response time properties that guarantee a quality of the service. The proposed approach gives readable models and an appropriate formal method which allows verification of required properties (no_deadlock, liveness, fairness) to prove the correctness of the workflow specification. In this context, several solutions have been proposed. Some of them use model checking for the verification. Van der Aalst [10] proposed a technique which uses Petri nets for the verification of the correctness of workflow applications using a compositional verification approach. Karamanolis et al [11] use process algebra for the verification of workflow properties. We have chosen to use the event B method and its associate refinement process and tools for the formal verification of workflow applications. The verification is based on a proof technique and therefore it does not suffer from the state number explosion occurring in classical model checking as in the cases of works in [10] and [11], which propose approaches for the verification of the correctness of the workflow specification.

In our previous work [2], we have proposed an approach which combines the use of UML AD and Event B for the specification and the verification of workflow applications. In this paper, we extend our work presented in [2] [13] by adding new translation rules for the UML AD endowed with timed characteristics and synchronisation aspects into Event B. Also, as in general the business domain depends on time consideration in their functionalities, we propose in this paper a solution to specify time in the event B method and derivation of temporal expressions in UML AD (timeout) into Event B. The result of the translation allows verification of the workflow termination. These translation rules give not only a syntactical translation, but also give a formal semantics using the Event B method semantics for the activity diagrams. In this context, there have been efforts for defining semantics for activity diagram in the works of Eshuis [7][12]. However, these works not consider the hierarchical decomposition of activities in UML AD, and suffer from the state number explosion. Moreover, in Eshuis [12][7] approach, no details are given about how time is defined. This paper is structured as follows. Section 2 describes the translation of hierarchical decomposition of UML AD into a hierarchy of Event B models. It presents derivation rules of time in UML AD into Event B notation. By an example we illustrate our contribution. Finally, a summary of our work concludes the paper.

II. THE PROPOSED APPROACH

As shown in Figure 1, the proposed approach consists mainly of four steps. In the first step, the workflow is modeled graphically with UML AD. In the second step, the resulting graphical readable model is translated into Event B in incremental development with successive refinements. This refined model is enriched by relevant properties (no deadlock, no livelock, strong fairness, etc) (Step3) which will be proved using the *B4free* tool [6] (step4). The verification of these properties ensures the correctness and the validation of the described workflow. In [2][13], we have proposed translation rules for the concepts of UML AD (activity, Sequence of activities, choice (decision), loop, parallel activities (fork and join), atomic process, and dynamic invocation) into Event B. Also, we have proposed, in [8], translation rules of event, send/receive event action concepts in UML AD into Event B. Due to space limitation we will not present all the proposed rules but just our proposed solution to present time in Event B, and the translation rules for the timeout concept in UML AD into Event B which will be illustrated over the ATM Login application example.

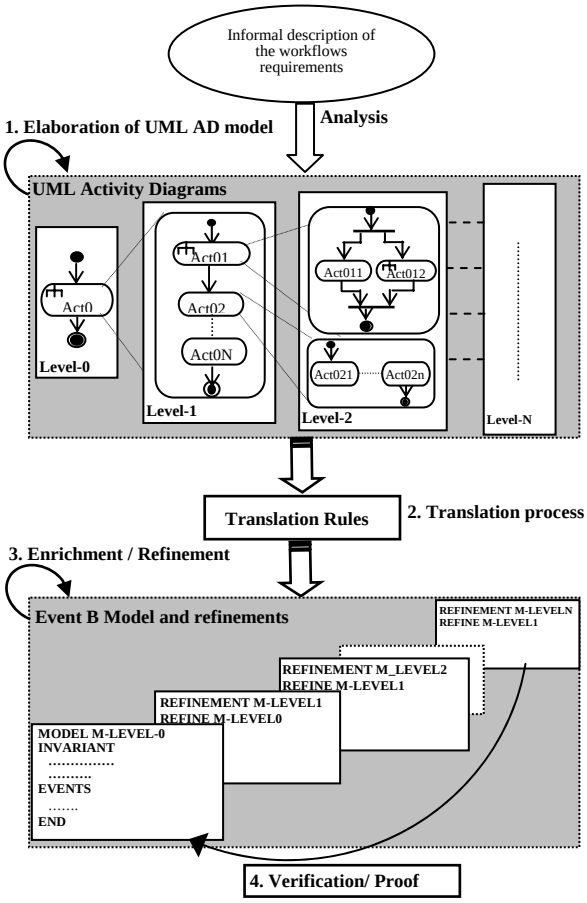


Figure 1. Derivation process from UML AD to Event B

Step 1: Specification with UML AD

Initially, we describe this workflow application using UML AD by employing a refinement technique [2]. The resulting model is composed of three decomposition levels (See Figure 2). Each activity has only one id (ATM, Card_Details, Eject_Card, User_PIN, Check_PIN, Get_out, Select_Transition, Get_Pin, Valid_PIN, Abounding, and End_Time).

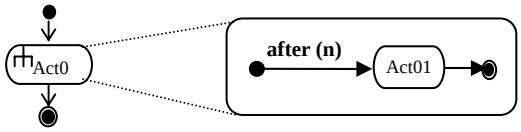
Step 2: Translation into Event B. Our proposed translation process uses the refinement process of Event B to encode the hierarchical decomposition in the UML AD: to each decomposition level in UML AD, is associated an Event B refinement.

The representation of the time in Event B

In UML, the time is incremented by means of an internal clock. The time is not a primitive of Event B. Specifying time in B, needs to add a clock in alternation with the system. We propose to model the action of this clock in Event B by the definition of a B event *Tick*. The *Tick*'s action consists of advancing the time represented by a B variable *Time* of integer type that we define in the build B system. The B event *Tick* maintains the control and allows time advance. In this way, we avoid the *Livelock* problem in the construction of a resulting B system. The timeout expressed in B, will impose alternation between the clock, the control system and the detection system. We use the variable *hand* [8] and the control is given alternatively to the clock when *hand=2* and to the system in the other cases.

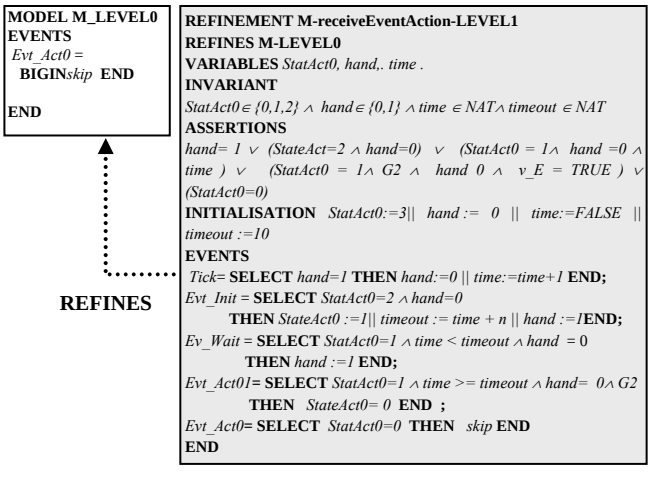
In UML AD, temporal event is specified with the after keyword. The after(*n*) event expression, where n is a positive integer, means that n time units after the source of the edge was entered a timeout is generated. The following figure shows an example.

Translation Rule of the event Timeout



The formulation of the timeout in Event B is based on some derivation rules that we already introduced. We propose to drift the timeout by:

- the definition of a integer variable *Timeout* which represents the time of the next generation of the event timeout;
- the definition of the event *Evt_Init* for initializing the variable *Timeout* with the value of the current time and the duration n (Exemple n =2 time unit) .
- the definition of the B event *Evt_Wait*.



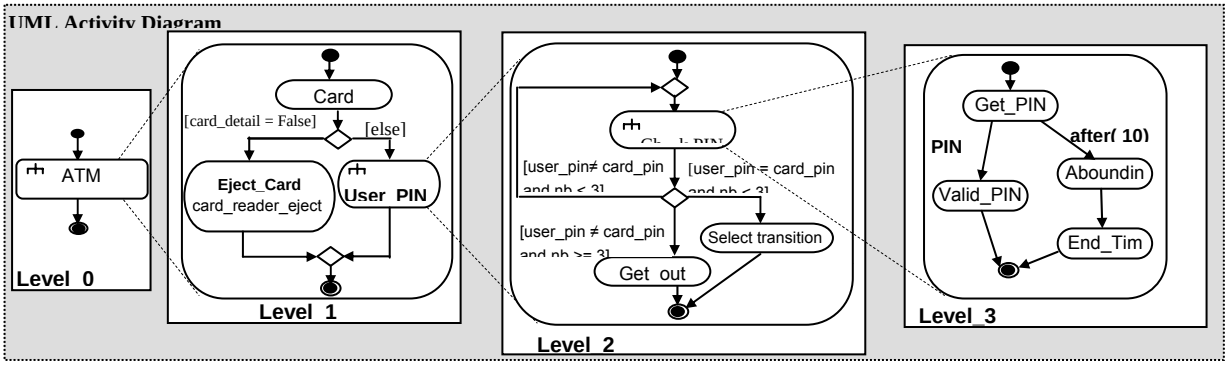


Figure2. The UML AD model of the workflow application ATM

Initially, *hand=0*. The event *Evt_Init* sets this variable to 1, and initializes the variable *timeout*. This passes the control to *Evt_wait or Evt_Act01*. If (*time < timeout)*) then the variable *hand* passes to 1 (event *Evt_Wait*) to increment the time (event *tick*) but if (*time>= timeout*) this (*Evt_Act01*) is allows the following activity Act01 to be execute.

By the application of the translation process and using the translation rules [2][13], the initial UML AD model of the ATM Login application, in figure 2, is translated into B event in a set of property preserving refinements. Three refinement steps which correspond to each level of three level of decomposition in the UML AD model (Figure 2) are necessary. Following figure2, the activities Card_Details, Eject_Card, Get_out, Select_transaction, Get_PIN, Valid_PIN, Abounding, and End_Time correspond to basic process/tasks of **ATM Login** application. We generate for the abstract level LEVEL0 an Event B model **ATMcard**, for the first decomposition level LEVEL1 an Event B refinement **Ref1_ATMcard**, for the second decomposition level LEVEL2 an Event B refinement **Ref2_ATMcard,** and for the third decomposition level LEVEL3 an Event B refinement **Ref3_ATMcard**. In following, we focus on the translation of event timeout.

In the refinement **Ref3_ATMcard**, the event *Tick* maintains the control and allows time advance. In this way, we avoid the *Livelock problem* in the construction of a resulting B system. By the application of the translation rule for the time, we use the variable *hand* and the control is given alternatively to the clock when *hand=2,* to the system when *hand= 0*, and to the detect system when *hand=1*. The variable *hand describes* the events

REFINEMENT Ref3_ATMcard
……..;
VARIABLES *hand, time, evt_pin, pin_state, timout …..*
INVARIANTS *hand ∈ {0,1,2} ∧ evt_pin ∈ BOOL∧ time ∈ N∧ timeout∈ N*
∧ pin_state∈{0,1,2,3 } */* the state variable pin_state is associated to the composed activity Chek_PIN*/*
INITIALISATIOIN *hand:= 0 || evt_pin:=FALSE|| time :=0|| pin_state:=3……*
EVENTS
Dect_Evt= **SELECT** hand =1 **THEN** hand := 0 || evt_pin :∈ BOOL **END;**
Tick= **SELECT** hand =2 ∧ pin_state=2 **THEN** hand :=1|| time := time+1 **END**;
Evt_GetPin= **SELECT** hand=0∧ pin_state=3 ….
THEN hand :=1|| pin_state :=2|| timeout := time+ 10 **END**;
Evt_Wait_Pin = **SELECT** hand=0∧ pin_state=2 ∧ evt_pin = FALSE ∧ time < timeout…
THEN hand := 2 **END**;
Evt_ValidPin= **SELECT** hand=0∧ pin_state=2 ∧ evt_pin = TRUE ∧ time < timeout
THEN pin_state :=0 **END**;
Evt_Abounding= **SELECT** hand=0∧ pin_state=2 ∧ evt_pin = FALSE ∧ time >=timeout
THEN pin_state := 1 **END**;

interleave and prevent that an event is fired infinitely (an event will be infinitely crossed in detriment of others).

Step 3: Fill up the system with properties

In this step, we enrich the models with invariants/Assertions describing required properties. The **ASSERTIONS** clause contains liveness properties expressing that there is ***no deadlock.*** This property is ensured by asserting that the disjunction of all the abstract events guards implies the disjunction of all the concrete events guards. This guaranties that the new events can be fired *(no deadlock*). In each new refinement, we add this property: for example in the **Ref2_ATMcard:**

REFINEMENT Ref2_ATMcard……..
ASSERTIONS
/* The disjunction of all the abstract events guards implies */
atm_state = 0 ∨ atm_state = 2 ∨ (atm_state = 1 ∧ card_detail = TRUE)
∨ (atm_state = 1∧ card_detail = FALSE)
=> /*the disjunction of all the concrete events guards*/
atm_state = 0 ∨ atm_state = 2
∨ (atm_state = 1 ∧ card_detail = TRUE∧ get_state= 1∧ user_pin ≠ card_pin∧nb > 0)
∨ (atm_state = 1 ∧ card_detail = TRUE∧ get_state= 1∧ user_pin = card_pin∧ nb > 0)
∨ (atm_state = 1 ∧ card_detail = TRUE∧ get_state= 1∧ user_pin ≠ card_pin∧ nb < = 0)
∨(atm_state = 1 ∧ card_detail = TRUE∧ get_state= 0)
∨ (atm_state = 1∧ card_detail = FALSE)

The **INVARIANT** clause allows to express the safety properties (called safety invariant) and the typing information (Typing invariant). Each refined model is enriched by relevant properties (safety, liveness, ect) which will be proved using the *B4free* tool. These properties shall remain true in the whole model and in further refinements: It is not needed to re-prove again verified properties in the refined model while the model complexity increases. It is the advantage of using *B4free* tool. For example:

- The safety property that the system ejects the card reader only if the card details are false: this property is added in the resulting refined model **Ref1_ATMcard** (associated to the **LEVEL1**) in the clause INVARIANT as follows:

REFINEMENT Ref1_ATMcard
REFINES ATMcard
INVARIANT */* Safety properties*/*
(*card_reader_eject = TRUE => card_detail= FALSE*)

- The temporal property T1 (the system should not be continuously open for more than 10 seconds without the even PIN present) can be proved by adding the safety invariant, in the resulting refined model **Ref3_ATMcard** (associated to the **LEVEL3** in UML AD model), which expresses

that if the system is in the node Abounding (*pin_state =1*), then necessarily the deadline has arrived:

REFINEMENT Ref3_ATMcard …….
REFINE Ref2_ATMcard
INVARIANT */*Temporal properties*/*
(pin_state = 1) ⇒ (time >= timeout) /* T1*/

The strong fairness (no livelock) properties are expressed by the events interleave by using the variable *hand* that solve livelock problem. The guards of these events define **their firing order** and how **these events interleave**.

Step 4: Validation/ Verification

A ATM subactivity is described by an initial state and a final state. It is refined into a sequence of basic events which lead from the initial state to the final one. The refinement preserves all the properties of the initial activity. This process is repeated until basic events are reached. When the basic events are reached by the refinement, the validation process is completed. The validation of the ATM activity allows to express reachability properties. For example, if a ATM activity is validated, then the objective of the application which consists of executing activity Card_Details then activity User_PIN *or* Eject_Card is realisable. In addition, for example this allows to express that while the event PIN is not detected in 10 s , after the execution of the activity Get_PIN, the activity Valid_Pin can not be executed. The following table1 illustrates the obtained results on our case study **ATM login**.

Model	nObv	nOp	nAuto	nInt	%Pr
ATM	0	0	0	0	100%
Ref1_ATMcard	32	4	4	0	100%
Ref2_ATMcard	84	10	10	0	100%
Ref3_ATMcard	255	25	25	0	100%
TOLAL	371	39	39	0	100%

Table 1. Summary of proofs, all Proof Obligations generated have been proved (Pr= 100%)

The resulting Event B specification has been proved totally. All proof obligations (100%) are proved automatically, and then the initial UML AD model of our **ATM** workflow application is validated.

III. CONCLUSION

In this paper, we have presented a formal syntax and semantic for UML AD endowed with time aspect and synchronisation (send/receive event). A systematic way for translating this semantics into Event B notation is also provided. Such a translation is not an end in itself, it is a basis for a formal and automatic verification of wide range of constrains including lives, no_deadlock , and safety properties , with the B support tool *B4free*, that ensure a confidence level for workflows applications such as business process. Currently, we are working on the implementation of this approach. Another thing needed to be mentioned is that we just formalize the subset of UML activity diagrams. For instance, object flows do not be included in our model. However, our approach is also suitable for formalizing it.

REFERENCES

[1] J.R, I. Jacobson, and G.Booch. “The Unified Modelling Language reference Manual” .Addison- Wesley, 1998.

[2] A. Ben Younes and L Jemni. Ben Ayed “ Using UML Activity Diagrams and Event B for Distributed and Parallel Applications”. In 31st Annual IEEE International Computer Software and Applications Conference (COMPSAC 2007), 24-27 July 2007, Beijing, China, Volume 1,

[3] M. Dumas and A.H.M ter Hofstede. “ UML activity diagrams as a Workflows Specification language ” . In *UML2001* page 76-90. Spinger-Verlag,2001.

[4] Clearsy. System Engineering Atelier B, Version 3.6, 2001.

[5] J.R. Abrial. “The B Book. Assigning Programs to Meanings”. Cambridge University Press, 1996.

[6] JClearsy, “B4free,” Available at http://www.b4free.com, 2004.

[7] R. Eshuis, R. Wieringa. *A formal semantics for UML Activity Diagrams – Formalising workflow models,* Technical Report. Twente, Dept. Of Computer Science,2001.

[8] A. Ben Younes and L.Jemni Ben Ayed ‘’ UML_AD2EventB: An Approach to Generating Event B Specification from UML Activity Diagrams for The Workflows Specification and Verification’’. In The third IEEE International Workshop on Scientific Workflows (SWF 42009). Los Angeles, California, USA, July 6-10, 2009.(Accepted)

[9] J-R Abrial.” Extending B without changing it” (for developing distributed systems)”. In H Habrias, editor, First B Conference, Nantes, France, 1996.

[10] W.M.P. van der Aalst, “Workflow Verification: Finding Control-Flow Errors Using Petri-Net-Based Techniques”, in *Business process management: models, techniques, and empirical studies.* LNCS 1806, Springer-Verlag, 2000.

[11] C. Karamanolis, D. Giannakopoulou, J. Magee, and S. M.Wheater, “Formal verification of workflow schemas,” University of Newcastle, Tech. Rep., 2000.

[12] R. Eshui, R, Wieringa. Tool Support for verifying UML Activity Diagram, IEEE transaction on software Engineering , vol 30 , N°7,juillet 2004

[13] A. Ben Younes and L Jemni. Ben Ayed “From UML Activity Diagrams to Event B for the Specification and the Verification of Workflow Applications”. In 32st Annual IEEE International Computer Software and Applications Conference (COMPSAC 2008), July 2008

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 500-503

A Web Services Composition Model for QoS Global Optimization

Minghui Wu[1,2], Xianghui Xiong[1,2], Jing Ying[1,2], Canghong Jin[2], and Chunyan Yu[3]
[1] Department of Computer Science and Engineering, Zhejiang University City College, Hangzhou, P.R. China
[2] College of Computer Science and Technology, Zhejiang University, Hangzhou, P.R. China
[3]College of Mathematics and Computer Science, Fuzhou University, P.R. China
Email: {minghuiwu}@cs.zju.edu.cn

***Abstract*—QoS driven selection approaches for Web Services Composition are used to choose the best solution among candidate services which have the same functions. This paper focuses on the Web services composition problem, and introduces the multi-dimension QoS model. Based on the QoS model, the QoS-driven web service selection model was proposed and described in details. According to these models, service composition problem can be considered as single-objective multi-constraints optimization problem. A brief discussion of approaches to solve the optimization problem is given.**

***Index Terms*—SOA; web services composition; QoS; global optimization;**

I. INTRODUCTION

Web services distributed in various locations can be integrated into a composite service with more powerful function. Through services composition, resources could be reused and we could implement a complicated functionality rapidly at lower cost.

A composite service is assembled by several tasks to accomplish a mission. In internet there are maybe many available web services with various QoS (Quality of Service) providing the same functionality specific to a task. So a selection needs to be made. More about QoS, you can refer to [4]. During the composition, there are demands for QoS constraints to be met and QoS criterions to optimize. Therefore web service composition has to search for an optimal set of services to construct a composite service and result in a best QoS, under user's QoS constraint and basic functionality claim. And how to construct the web service composition model is the main research subject of the paper.

The remainder of this paper is organized as follows: the section 2 provides some related work of web services composition, and the section 3 gives the web services composition model. A particular depiction of the QoS-driven web service selection mathematical model and algorithm has been given in section 4. Finally, the last section concludes this paper and prospects the future.

II. RELATED WORKS

Some approaches of web service selection are based on semantic web [2, 3], others are basis of QoS attribute computing [4-6]. The former has difficulties in global QoS evaluation. At present there are many approaches base on QoS attribute computing. Zeng [4] proposes two ways, local optimization and global planning by using integer programming. Local optimization will obtain optimal candidate services in one abstract service scope without considering constraints across multiple abstract services and contributions to optimize global QoS criterions. Consequently, the binding of abstract services is independent with each other. Unless QoS of service is changed or unavailable, the service binding do not need to redo. The time consumed will be stable if we keep identical scale of service composition. However, unlike global planning which can get better global QoS, local optimization could not follow global criterion. Nevertheless, global optimization also has its defect: the time cost will grow proportionally with the amount of execution routes for the same scale of services. Yu [6] defines the service composition problem as a Multi-dimensions Multi-choice 0-1 knapsack problem (MMKP) in combinatorial model and a Multi-constraint optimal path (MOP) problem in graph model, and solves it by using integer programming. Li [7] proposes a mapping framework to construct Service Overlay Network (SON), and translates web service composition problem to single constraint path selection problem. In the end, Dijistra shortest path algorithm could give an optimal selection.

We will construct the web service composition model and demonstrate its mathematical model as single-objective multi-constraints optimization problem.

III. WEB SERVICES COMPOSITION MODEL

A. Basic Definition

Definition 1: (Abstract Service). Abstract service has function descriptions without implementation and standard service interfaces across different service providers. An abstract service is corresponding to a workflow task.

Definition 2: (Service Instance). Service instances are concrete services published by service providers. They could give the function implementation specified by abstract services. And some service instances may have the same function, but different QoS.

Definition 3: (Candidate Relationship). While the function of several service instances $S_1,...,S_n$ are consisted with the function description of abstract service T, we state that service instances $S_1,...,S_n$ and abstract service T have the candidate relationship. We also say

that $S_1,...,S_n$ is the candidate services of T which is labeled as $S_i \in T \quad (i=1...n)$.

Definition 4: (Service Function Graph). Constructing abstract services as a workflow to fulfill user's requirement in functionality will obtain a service function graph. In the service function graph, we have two additional special abstract services treated as Start label and End label, which have no functional meaning. An example is given as Fig.2.

Definition 5: (Service Selection Graph). Shown as Fig.3, all the service instances corresponding to abstract services in Service Function Graph of Fig.2 have been discovered, and a service function graph appears. In the process of service discovering, we will discard service instances that do not meet local constraints. Consequently, local constraints are already met in service selection graph, and we will not take account of local constraints in the following discussion.

Definition 6: (Execution Route). In the service function graph, execution route is a passageway between start node and end node, and only include one spur track for every condition structure. If there are k execution routes with a probability $\rho_i (i=1...k)$ in a service function graph, then $\sum_{i=1}^{k} \rho_i = 1$.

In Fig.2, there are two execution routes:

$ER_1 : (S_1, S_2, S_3, S_4, S_6)$, with probability ρ_1

$ER_2 : (S_1, S_2, S_3, S_5, S_6)$, with probability ρ_2

and $\rho_1 + \rho_2 = 1$.

Definition 7: (Execution Plan). For execution route $(S_1,...,S_n)$, we define $(T_1,...,T_n)$ as a execution plan of execution route $(S_1,...,S_n)$, if $T_i \in S_i \quad (i=1...n)$, here T_i denotes service instance, S_i represents abstract service. According to the definition, an execution plan is an executable service chain which meets user's requirement.

In Fig.3, execution plans of execution route ER_1 could be: $(S_{11}, S_{21}, S_{31}, S_{41}, S_{61}),...,(S_{12}, S_{23}, S_{33}, S_{42}, S_{63})$.

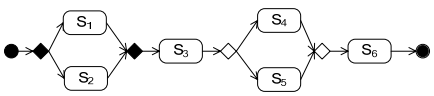


Figure2. Service function graph

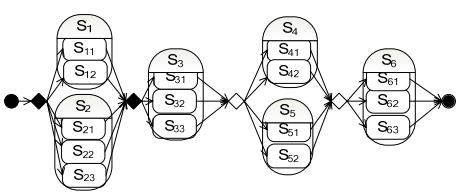


Figure.3 Service selection graph

B. Web Service Composition Flow

We abstract web service composition flow as three phases:

Firstly, compose abstract services as a service function graph to meet user's function demand in abstract hierarchy.

Second, discover all the functionality demand in abstract hierarchy. Service instances which have candidate relationships with abstract services. And transform service function graph to service selection graph.

Last, select a service chain linking the start node and end node in service selection graph, and the chain could keep the QoS constraints.

C. Multi-dimensions QoS Model

QoS is a series of non-function attributes. In this paper we will just use three attributes: reputation, price and availability into account. Other attributes could be imported and will not impact our results. Due to article [4] has already given an explicit depiction of definition and calculation of QoS attribute, we will not give a duplicate description.

C.1 QoS Attribute Standardization and Utility Function

Different QoS values have different value ranges, it is unfair to calculate these values directly. So QoS attribute standardization is needed. And in order to evaluate quality of the overall QoS of service, utility function is imported.

In this paper, we states the QoS of the jth service instances S_{ij} included in abstract service S_i is $(q_{j1},...,q_{jn})$. QoS attributes can be separated into positive attributes and negative attributes. The positive attribute, as availability and reputation, means the higher value the better selection result. Negative attribute, as price and duration time, is contrast with positive one. So we standardize positive attribute with formula (1) and standardize negative attribute with formula (2)

$$q'_{jk} = \frac{q_{jk} - \alpha_k}{\sigma_k} \quad 1 \le k \le n \tag{1}$$

$$q'_{jk} = 1 - \frac{q_{jk} - \alpha_k}{\sigma_k} \quad 1 \le k \le n \tag{2}$$

Where α_k, σ_k are the average and standard deviation of the QoS values for all candidates of kth QoS attribute of service.

A service has multi-dimensions QoS value $(q'_1,...,q'_n)$, so we need to consider all the QoS in web service selection process. But multi-dimensions value is not easy for comparison. Thus we propose a utility function, mapping multi-dimensions value into a function value which is a scalar quantity, to give a comprehensive reference for our comparison. The definition of utility function is as follows:

$$UF(s) = \sum_{i=1}^{n} \omega_i \cdot q'_i \quad \text{and} \quad \sum_{i=1}^{n} \omega_i = 1 \tag{3}$$

Where ω_i denotes the weight of the ith QoS attribute which reflects their importance. q'_i is the standard attribute value.

C.2 QoS of Execution Plan (Composite Service)

Execution plan $EP(S_1,\ldots,S_n)$ is a powerful service composed by a series of atomic services. Like a normal service, execution plan has its own QoS. The QoS calculation formula of execution plan is displayed in Table 1.

TABLE I.
QOS CALCULATION FORMULAS

QoS attribute	Formula
Price	$q_{price} = \sum_{i=1}^{n} q_{price}^{'}(i)$
Reputation	$q_{reputation} = \frac{1}{n}\sum_{i=1}^{n} q_{reputation}^{'}(i)$
Availability	$q_{availability} = \prod_{i=1}^{n} q_{availability}^{'}(i)$
Others	Formulas decided by attribute

C.3 QoS of Execution Route

In execution route $ER(S_1,\ldots,S_n)$, the candidates of abstract service $S_i(1 \le i \le n)$ are $S_{i1},\ldots,S_{il}$, and QoS of the jth service instance S_{ij} is $(q_{j1}^{'},\ldots,q_{jm}^{'})$, so the QoS calculation formulas are as follows:

$$Q^{repu} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{i_l} x_{ij} \cdot q_{ij}^{repu}}{n} \quad (4)$$

$$Q^{price} = \sum_{i=1}^{n}\sum_{j=1}^{i_l} x_{ij} \cdot q_{ij}^{price} \quad (5)$$

$$Q^{reliability} = \prod_{i=1}^{n}\sum_{j=1}^{i_l} x_{ij} \cdot q_{ij}^{reliability} \quad (6)$$

When the service jth is selected as an instance of the abstract service ith, we set $x_{ij}=1$, otherwise we set $x_{ij}=0$. The i_l denotes the amount of candidate services of the abstract service i. Other QoS calculation formula of execution route could be imported by the same way.

Ⅳ. QOS-DRIVEN WEB SERVICE SELECTION MATHEMATICAL MODEL AND ALGORITHM

A. QoS-driven Web Service Selection mathematical Model

QoS-driven web service selection problem could be mapped into single constraint multi-objectives optimization problem in mathematics. We will first give web service selection mathematical model for single execution route, then web service selection mathematical model for multi execution routes will be displayed.

A.1 QoS-driven Web Service Selection Mathematical Model for Single Execution Route

An execution route could make multi execution plans when abstract service is bound to different candidate services. So web service selection problem has to choose the optimal execution plan among all execution plans. In mathematics, it could map into a single constraint multi objectives optimization problem.

First of all, topological sequence all the abstract services in an execution route, then each abstract service has its own id i. Also we do the same things to all candidate services of each abstract service, and every candidate has an id j. We assume that abstract service $S_i \ (1 \le i \le n)$ in execution route $ER(S_1,\ldots,S_n)$ has candidate services $(S_{i1},\ldots,S_{i_l})$, and i_l is the amount of service instances included in abstract service S_i. Then, the corresponding single-objective multi-constraints optimization problem is:

1) Objective function

$$MAX \quad f(ER) = UF(ER) = \sum_{i=1}^{n} \omega_i \cdot Q_i^{'} \quad (7)$$

Where $Q_i^{'}(1 \le i \le n)$ is the ith QoS attribute value of execution route ER, derives from QoS calculation formula of execution route for standardization QoS. ω_i denotes the weight.

2) Global QoS constraint

Regarding that QoS values in global QoS constraints are actual data in application, so all the QoS values are not standardized in this part. Global QoS constraints can be divided into two groups.

- Single selection constraint

In the web service selection process, there is only one candidate service of each abstract service can be selected into composition flow. So single selection constraint can formalize as below:

$$\forall i \in n \quad , \quad \sum_{j=1}^{i_l} x_{ij} = 1 \quad (8)$$

Where n denotes the number of abstract services in execution route. When jth candidate service in abstract service i is selected, $x_{ij}=1$; Otherwise, $x_{ij}=0$, i_l is the amount of service instances of abstract service i.

- QoS value constraint

QoS value constraints are constraints proposed by users, such as service price don't go beyond 100$, service availability don't be under 99.9%. If there are h QoS value constraints $C^k, 1 \le k \le h$, and k denotes the kth QoS attribute, then

$$\forall k, \ 1 \le k \le h \ ; \ Q_k \le C^k \quad (9)$$

Where Q_k is the QoS value with attribute k of execution route i.

A.2 QoS-driven Web Service Selection Mathematical Model for Service Function Graph

There are may be multiple execution routes in a service function graph. Web service selection of service function graph with multiple execution routes is more close to the application in reality. The problem can map into single-objective multi-constraints optimization problem in mathematics.

A.2.2 Web Service Selection Mathematical Model of Service Function Graph with Multiple Execution Routes

Abstract web service selection problem of multiple execution routes into single-objective multiple-constraints optimization problem.

The execution routes of service function graph are $ER_1, ER_2, \ldots, ER_s$, and execution probability of $ER_i \quad (1 \le i \le s)$ is ξ_i, where $\sum_{i=1}^{s} \xi_i = 1$.

1) Objective function

$$MAX \quad f(SFG) = \sum_{i=1}^{s} \xi_i \cdot f(ER_i) \tag{10}$$

Where $f(ER_i)$ is defined by formula (7).

2) Global QoS constraints

- Selection constraints

It is the same as formula (8).

- QoS value constraints

In order to meet the QoS value constraints strictly, we enforce every execution route to satisfy the constraint requirements. If there are h QoS value constraint $C^k, 1 \le k \le h$, and k denotes the kth QoS attribute, then

$$\forall i,k \quad 1 \le i \le s \quad 1 \le k \le h \quad ; \quad Q_k^i \le C^k \tag{11}$$

Where Q_k^i is the kth QoS attribute value of the execution route i.

B. QoS-driven Web Service Selection Algorithm

Integer programming and evolutionary algorithm, such as genetic algorithm, are two methods adopted broadly to solve optimization problem of web service selection. The strongpoint of QoS attribute computing based on integer programming lies on the maturation of theory and plenty of approaches. But it claims that the QoS constraints and QoS criterions should be linear while there are many non-linear QoS attributes just like availability. Genetic Algorithm also be applied to web service selection [5,8,9]. In contrast to Integer Programming, it has no requirement on whether the QoS constraints and objective function are linear or not, which extends the field of application rooting from Genetic Algorithm's dependence to problem areas. On the other hand, Genetic Algorithm has better time consumed than Integer Programming as the increasing of composition scale [8].

Ⅴ. Conclusion and Future Work

Our work at present mainly focuses on constructing a platform for SOA application development: SMICE (Semantic Model-driven Integrated Construction Environment). Research of service composition model is an important part in this project. At the early stage we had developed a QoS driven service composition ontology framework [10], which provides a basis for web service composition and selection. Our work in this paper mainly lays on giving a QoS-driven web service composition model in detail and map it into single-objective multi-constraints problem. In future how to combine service semantics and QoS attribute to integrate more efficiently is an important goal of our work.

Acknowledgment

This work was supported in part by the National High-Tech Research and Development Plan Foundation (863), China (Grant No.2007AA01Z187) and NSFC (Grant No. 60805042)

References

[1] J. Cardoso, *Quality of Service and Semantic Composition of Workflows*, University of Georgia, 2002.

[2] C. Zhou, L.T. Chiq, B.S. Lee, "DAML-QoS Ontology for Web Services", in *Proceedings of the 2004 IEEE International Conference on Web Services* (ICWS2004), 2004, pp. 472-479.

[3] A.S. Bilgin, M.P. Singh, "A DAML-Based Repository for QoS-Aware Semantic Web Service Selection", in *Proceedings of the 2004 IEEE International Conference on Web Services* (ICWS2004), 2004, pp. 368~375.

[4] L.Z. Zeng, B. Benatallah, A.H.H Ngu et al, "QoS-Aware Middleware for Web Services Composition", *IEEE Trans. Softw. Eng.*, Vol.30, No.5, 2004, pp. 311-327.

[5] M. D. P. G Canfora, R Esposito, M.L Villiani, "A Lightweight Approach for QoS-Aware Service Composition", in *Proceeding of 2nd International Conference on Service Oriented Computing*, 2004, pp. 36~47.

[6] T. Yu, Y. Zhang, K.J. Lin, "Efficient algorithms for Web services selection with end-to-end QoS constraints", *ACM Trans. Web*, Vol.1, No.1, 2007.

[7] Y. Li, J. Huai, T. Deng et al, "QoS-aware Service Composition in Service Overlay Networks", in *Proceeding of the 2007 IEEE International Conference on Web Service* (ICWS2007), 2007, pp. 703~710.

[8] G. Canfora, M.D. Penta, R. Esposito et al, "An approach for QoS-aware service composition based on genetic algorithms", in *Proceedings of the 2005 Conference on Genetic and evolutionary Computation*, 2005, pp. 1069~1075.

[9] L.J. Zhang, B. Li, T. Chao et al, "On demand Web services-based business process composition", in *Proceeding of IEEE International Conference on System, Man and Cybernetics* (SMC'03), vol.4, 2003, pp. 4057~4064.

[10] M.H. Wu, C.H. Jin, C.Y. Yu, et al, "QoS and Situation Aware Ontology Framework for Dynamic Web Services Composition", in *Proceeding of Proceedings of the 2008 12th International Conference on Computer Supported Cooperative Work in Design* (CSCWD' 08), Vol.1, 2008, pp. 459~464.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 504-506

Advanced Dynamic Source Routing with QoS Guarantee

Youyuan Liu
School of Computer Science, Chongqing University of Arts and Sciences, Chongqing, China
cqyclyy@163.com

***Abstract*—With a great deal of different applications increasing, it needs to provide effect communication for mobile Ad Hoc network. In this paper, we present an advanced dynamic source routing (ADSR) algorithm for such network. ADSR can select routes according to link state and dynamic delay detection. In the route discovery phase, ADSR finds paths with the greatest link stability factor. In the route maintenance phase, it effectively keeps monitoring network topology changes by delay prediction and timely performs rerouting before the paths become unavailable. With such enhancement strategies, ADSR can significantly improve the overall performance of network. Experimental results show that ADSR can achieve higher packet delivery rate and at the same time retain lower end-to-end delay comparing with conventional DSR algorithm.**

***Index Terms*—routing algorithm, dynamic source routing, MANET, QoS**

I. INTRODUCTION

Mobile Ad Hoc network (MANET) is a multi-hop temporary autonomous system of mobile nodes with wireless transmitters and receivers without the aid of pre-established network infrastructure [1]. Such network is created spontaneously without any infrastructure. The placement of nodes, in most of the cases, is dependent upon the application and is unpredictable. The nodes are not managed or controlled by any central node. Communication among nodes, in this type of networks, can be either direct or via relaying nodes. Since the network can also dynamically change its topology, routing has a crucial impact on the network performance. Moreover, different services in MANET have raised certain research concerning routing with Quality of Service (QoS) supporting. The QoS requirement is defined as a set of constraints to be met by a network while in communication on performance metrics, such as bandwidth, delay or link state. Routing algorithms with different metrics have been proposed in many literatures [2][3][4][5].

In [6], Goff and Ghazaleh investigate to add proactive route selection and maintenance for on-demand MANET routing algorithms. When a path is likely to be broken, a warning is sent to the source indicating the likelihood of a disconnection. The source can then initiate path discovery early, potentially avoiding the disconnection altogether. In [7], Queuing theory and communication theory are jointly applied to relate a routing metric called permissible arrival rate, which can support routing discovery under an average delay constraint. W. Zhua proposes a novel algorithm called ticket-based probing with stability estimation (TBP-SE). Models are created to estimate relative link and path stability [8][9]. In the context of geographic (location-based) routing, a scheme is proposed in to predict future paths before existing paths break [10]. This scheme can avoid path re-computation delay. But it does not reduce path breakage so that some problems such as transmission failure and huge routing message overhead still exist. In [11], Wang et al. respectively define link stable time for a link and path stable tim for a path. If the stable time of a path is going to be expired, the source node will discover a new path in advance. However, no detailed method is given to estimate the link stable time. In [12], analysis and simulations show a significant network capacity gain for MANET employing multiuser detectors, compared with those using matched filter receivers, as well as very good performance even under tight delay constraints. Xie et al. propose a link reliability based hybrid routing, which is a novel hybrid protocol. Contrary to the traditional single path routing strategy, multiple paths are established between a pair of source-destination nodes. In the hybrid routing strategy, the rate of topological change provides a natural mechanism for switching dynamically between table-driven and on-demand routing [13]. A. Kherani and R. El-Khoury study the throughput of multi-hop routes and stability of forwarding queues in MANET with random access channel. Their result is characterization of stability condition and the end-to-end throughput using the balance rate. They show that as long as the intermediate queues in the network are stable, the end-to-end throughput of a connection does not depend on the load on the intermediate nodes. In this paper, we propose an advanced DSR algorithm which can significantly improve the overall performance of network

II. PARAMETER CALCULATION

We can represent MANET by a weighed graph $G(V, E)$ where V is the set of nodes in the network and E is the set of links with connected nodes which are in transmission range of each other. Since V and E change with the moving, joining and leaving of nodes, MANET has a dynamic topology. Two mobile nodes within transmission range are not enough to ensure the successfully communication, since many phenomena, such as interference, physical obstacles and power problems, may occur during the transmission and cause it to fail. By using intermediary nodes to forward the message can tackle with the problem.

In MANET, the difference of delay is often great. We introduce a simple method similar to traditional RTT adaptive algorithm to evaluate the delay variety. Assume that D_{cur} denotes the current delay of routing, D_{old} denotes the previous delay of routing and D_{new} denotes the future delay predicted. If ΔD_{cur} is the current delay variety, $\Delta D_{cur} = D_{cur} - D_{old}$. Thus, $\Delta D_{new} = a \cdot \Delta D_{old} + (1 - a) \cdot \Delta D_{cur}$ and $D_{new} = D_{cur} \pm \Delta D_{new}$. The equations can be used to predict the dynamic delay. Since the topology in MANET often changes, the delay is also dynamic changes. Thus the bound of delay can be considered. Assume that D_{req} is the delay constraint for QoS request. We can compare D_{req} with $b \times D_{new}$, where b is the bound factor which will be set to 1.2 empirically.

Once the distance between two mobile nodes exceed a certain threshold, it may result in link failure and need routing rediscovering, which will inevitably decrease routing delay and packet loss ratio. In order to solve this problem, a link stability measure can be introduced into routing algorithm which is capable of predicting the duration of time routes will remain valid. For two neighbor nodes i and j, assume that their coordinates are (x_i, y_i) and (x_j, y_j) while velocity are v_i and v_j. Furthermore, the velocity decomposition along X-axis and Y-axis of the two nodes are (v_{ix}, v_{iy}) and (v_{jx}, v_{jy}) respectively. While exceeding a period t, the coordinate of the two nodes are $(x_i + t\, v_{ix}, y_i + t\, v_{iy})$ and $(x_j + t\, v_{jx}, y_j + t\, v_{jy})$. Finally, we have $t = \dfrac{\sqrt{r^2(a^2+b^2) - (ac-bd)^2} - (ad+bc)}{a^2+b^2}$. Note that $a = v_{jx} - v_{ix}, b = v_{jy} - v_{ij}, c = y_j - y_i, d = x_j - x_i$. t is defined as the link stability factor and it represents the time needed for two nodes reaching the maximum effective transmission distance. Once exceeding such period, the current route will fail. Routing algorithm considering about such factor will achieve greater link stability than traditional routing algorithm.

III. Our Proposed Routing Algorithm

In on-demand protocols such as DSR, nodes only compute routes when they are needed. Like any source routing protocol, in DSR the source includes the full route in the packets' header. In this paper, we present an advanced dynamic source routing (ADSR) algorithm which extends the DSR protocol by flooding RREQ to build routes. To utilize the information obtained from the mobile prediction scheme, extra fields must be added into conventional RREQ in DSR protocol. When a source node floods RREQ, it appends its location, speed, and direction into the control packet. It sets the maximal link expiration time to the corresponding field firstly. When the relay node receives a RREQ, it will predict the link expiration time between itself and the previous hop. The minimum between this value and the link expiration time recorded in the RREQ is included in the packet. Once a single link on a certain path is disconnected, the entire path is invalidated. The node need update the location and mobility information field written by the previous node with its own information. If a relay node receives multiple packets with different link expiration time, it selects the minimum value among them and sends its own routing table with the chosen link expiration time attached.

ADSR includes routing discovery and routing maintenance phases. Routing discovery process is to find feasible paths between source and destination node. Routing maintenance process is to monitor and predict the future information about availability of link. When a source node initiates a routing request to a destination node, it first checks its routing cache. If there exists feasible paths, the source selects the most stable one to send data packets. While a node receives a RREQ, the process is as follows:

(1) If it is not the destination, it will check its ID with those received before. Once exists same IDs, it means that the RREQ was repeatedly received and must be discarded.

(2) It begins to check whether the delay constraint is met. If not satisfied, the RREQ will be discarded.

(3) It begins to calculate the link stability factor. With the information of position and velocity of upstream node and itself, relay node calculates the link stability factor t.

(4) If $t <$ LET, it will be modified with t as the new LET. Otherwise, LET remains unchanged. Thus, all relay nodes along a route will stay the minimum stability factor.

(5) Information about position and velocity of RREQ will be added to the corresponding fields of the relay node. And their address will be added to the corresponding fields of RREQ. Then RREQ will be forwarded.

When the destination node receives RREQs, it will obtain all feasible paths meeting the delay requirement. It begins to send RREP to the source and the link stability time will be copied in RREP. When the source receives RREP, it calculates the end-to-end delay of each route and stores it to routing cache. Thus each routing cache includes corresponding link stability time and delay. Finally, the source will select the best route among all feasible paths and use it to send data packets. Once the source node receives a RRER, it will delete the corresponding routes from its own routing cache. Then it checks whether there are other routes to destination in the routing cache. Among all of the feasible routes found, it selects the best route to send packets.

IV. Experimental Results

In this section, we compare the performance of standard DSR with ADSR algorithm through simulation experiments with different network scenes. In our experiments, networks with a specified number of nodes are randomly generated within a 1000×1000 square region. One pair of the nodes is randomly chosen to be the source and destination. The IEEE 802.11 MAC protocol is used in the network. Random way-point is

selected as movement model and the CBR is used to send data.

In the experiments, we generate a network with 60 nodes, and the delay request is set to 40ms. The pause time of mobile nodes varies from 0 to 150s. As is shown in figure 1 and 2, when the pause time of nodes decreases, which means the network topology changes frequently, the performance of both algorithms degrade. Since the establishment of unstable routes increases, it will result in high overhead for routing reconstruction. In ADSR, however, those RREQs can not meet the delay requirement will be discarded and in turn reducing the number to reconstruct routing. Once the feasible route is established, the stability of data transmission will naturally increase because of its mobile prediction scheme. In the case of more static topology, the delay of ADSR is higher than that of DSR because it needs to calculate link stability factor, which will cause more latency in routing discovery process. However, the decrease of pause time of nodes will lead to more links disconnected. The routing delay can be offset in ADSR and it will get better performance because of its stability first routing scheme.

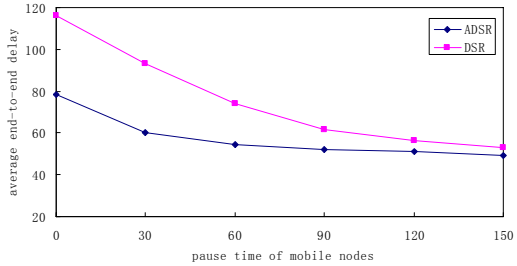


Fig.1. End-to-end delay vs. mobility

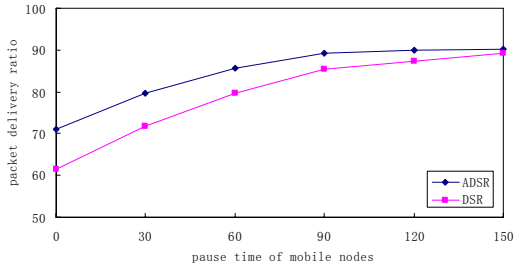


Fig.2. Packet delivery ratio vs. mobility

V. CONCLUSIONS

MANET is an autonomous system of mobile nodes with wireless transmitters and receivers without the aid of pre-established network infrastructure. In this paper, a new routing algorithm is proposed which can dynamically estimate delay variety and calculate link stability factor. In the route discovery phase, it finds paths meeting delay requirement with greatest stable link. In the route maintenance phase, it effectively keeps monitoring network topology changes through delay prediction and performs rerouting in time. Simulation results show that it can meet with the delay requirement while decrease the routing cost. Comparing with other conventional routing algorithms, it can achieve higher packet delivery rate in dynamic network.

ACKNOWLEDGMENTS

This work was supported by the Scientific Research Fund of Chongqing University of Arts and Sciences under grant Y2008SJ32.

REFERENCES

[1] J. G. Jetcheva and D. B. Johnson, “Routing characteristics of ad hoc networks with unidirectional links,” *Ad Hoc Networks*, Volume 4, Issue 3, 2006, pp. 303-325.

[2] H. Xiao and G. SEAH, “A Flexible Quality of Service Model for Mobile Ad-Hoc Networks,” *IEEE VTC2000-sprin*, Tokyo, 2000, pp. 2-7.

[3] M. Joa and T. Lu, “A Peer-to-Peer Zone-Based Two-Level Link State Routing for Mobile Ad Hoc Networks,” *IEEE Journal on Selected Areas in Communications, special issue on Wireless Ad Hoc Networks*, 1999, pp. 1415-1425.

[4] Xiaojiang Du, “QoS routing based on multi-class nodes for mobile ad hoc networks,” *Ad Hoc Networks*, Volume 2, Issue 3, 2004, pp. 241-254.

[5] S. B. Lee, “INSIGNIA: An IP-based quality of service framework for mobile Ad Hoc networks,” *Journal of Parallel and Dist. Comp., Special issue on Mobile Computing and Communications*, 2000, pp. 374-406.

[6] T. Goff, N. A. Ghazaleh, D. Phatak and R. Kahvecioglu, “Preemptive routing in ad hoc networks,” *Journal of Parallel and Distributed Computing*, Volume 63, Issue 2, 2003, pp. 123-140.

[7] V. Srivastava and M. Motani, “Combining Communi-cation and Queueing with Delay Constraints in Wireless Ad-Hoc Networks”, *ICICS-PCM 2003*, Singapore, 2003, pp. 1086- 1090.

[8] Son, A. Helmy, and B. Krishnamachari, “The Effect of Mobility-Induced Location Errors on Geographic Routing in Mobile Ad Hoc and Sensor Networks: Analysis and Improvement Using Mobility Prediction,” *IEEE Trans. on Mobile Computing*, vol. 3, no. 3, 2004, pp. 233-245.

[9] S. H. Shah and K. Nahrstedt, “Predictive Location-based QoS Routing in Mobile Ad Hoc Networks,” *Proc. of 2002 IEEE International Conference on Communi-cations* , 2002, pp. 1022-1027.

[10] W. Zhua, M. Songa and S. Olariub, “Integrating Stability Estimation into Quality of Service Routing in Mobile Ad-hoc Networks”, *14th IEEE International Workshop on Quality of Service*, IEEE, 2006, pp. 122-129.

[11] Y. Wang and S. Chang, “Interfering-aware QoS Multi path Routing for Ad Hoc Wireless Network,” *Proc. of 18th International Conference on Advanced Information Networking and Applications*, 2004, pp. 29-34.

[12] C. Comaniciu and H. Vincent Poor, “On the Capacity of Mobile Ad Hoc Networks with Delay Constraints”, *IEEE TRANSACTIONS ON WIRELESS COMMUNI-CATIONS,* VOL. 5, NO. 8, 2006, pp. 2061- 2071.

[13] Xie Xiaochuan, Wei Gang, Wu Keping, Wang Gang and Jia Shilou, “Link reliability based hybrid routing for tactical mobile ad hoc network,” *Journal of Systems Engineering and Electronics*, Volume 19, Issue 2, 2008, pp. 259-267.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 507-509

MATLAB Simulation of Paroxysmal Public Crisis Information Dissemination Based on the Network

Zhihong Li[1] , Guanggang Zhou[2] , and Xin Wei[3]
[1,2]School of Business Administration, South China University of Technology
Guangzhou, 510640, China
bmzhhli@scut.edu.cn
[3]School of Sciences, South China University of technology
Guangzhou, 510640, China
eeee-85@163.com

***Abstract*— The network has played an inestimable role in the process of the dissemination of information during the crisis as a new medium. The spread of crisis information on the state of the network depends largely on the amount of information to be found on the network. This paper used the classical single-cylinder model in probability theory to simulate the process of accumulation of crisis information. Through the MATLAB simulation could found in the dissemination of crisis information there was uncertainty. And Positive and negative information would emerge Matthew in the accumulation process. The government needs to advance to guard against the negative information appears on the Matthew Effect.**

***Index Terms*—Paroxysmal public crisis; Matthew Effect; Internet Communication; Single-cylinder model; MATLAB**

I. FOREWORD

From the Indian Ocean tsunami to H1N1 in Taiwan, from the 911 terrorist attacks to Tibet riots, seemingly peaceful society bears a variety of crisis. Positive messages can come together to fight crisis, But the negative information will result in rumors all over the city, people panic, and even a crisis caused by a variety of secondary. In particular, the rapid development of the Internet today, the spread of crisis break time and region, all kinds of information is likely to increase in geometric approach, so dissemination of crisis information has brought new challenges to control.

Unbalanced accumulation of different types of information often occur Matthew behavior. As the people's herd mentality, once a dominant point of view, its credibility is getting high, the possibility of being re-transmission also increased. Single-cylinder model polya distribution used probability theory to explain why there was Matthew. This paper will use the single-cylinder model to simulate the process of crisis information dissemination on the network based on MATLAB to explore the control and governance approaches for the government in a crisis.

II. MODEL CONSTRUCTION AND ANALYSIS

Polya distribution also known as single-cylinder model, it is one of the classic model of information accumulation. The model is based on the following test: Suppose a cylinder is equipped with a number of red balls and black balls, according to certain rules to take the ball from the jar, red ball means success and black means fail. Suppose there are A red balls and B black balls, and selected one ball arbitrarily. If it was a red ball, then put it and other C red balls in the jar, and black ball is also. Repeat this experiment. This article took communication process as the process of the ball to take place.

A. Models hypothesis

According to characteristics of the crisis information dissemination on network, a crisis communication model to make the following hypothesis.

a) For the complexity of the crisis information dissemination, this article did not consider the diversity of content and form of crisis information. Suppose crisis information can be divided into two categories, one is positive message, note the positive information, with a red ball means; the other is negative message, recorded as negative information, with a black ball means.

b) The crisis information dissemination is uncertainty, in the free dissemination of the process, suppose that for each message was selected for further propagation (the development of new information) of the same possibility.

c) Only consider the effective communication in the process of dissemination. And that information won't to be distorted; positive news does not spread into the negative message in the process of dissemination.

d) The same type of information at the same rate.

B. Single-cylinder model

According to test procedure of Single-cylinder model and repeated the experiment N times, then the number of red balls namely the number of positive messages in the jar is:

$$a_n = a_{n-1} + p_n c ,$$

Also the number of black ball namely the number of negative messages is:

$$b_n = b_{n-1} + q_n c ,$$

We can define p_n represents the probability of getting positive information, and q_n represents the probability of getting negative information. After taking the ball every time then put C or D the same color ball to show accumulation process of crisis information dissemination.

In the dissemination process, we not only care about the growth of red balls and black balls and also concerned about changes in the proportion of the ball. In the completely random state, given a set of initial value ($a = 30$, $b = 50$, $c = d = 10$) to carry out MATLAB simulation. Three continuous operation got different results. (Red indicating that the probability of red ball, blue indicating that the probability of the black balls)

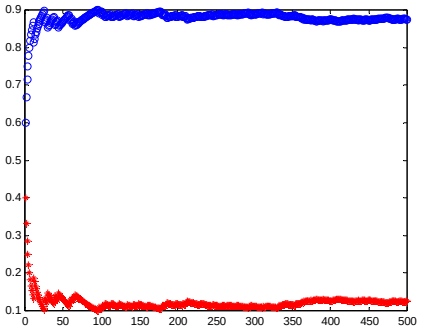


Fig 1(a)

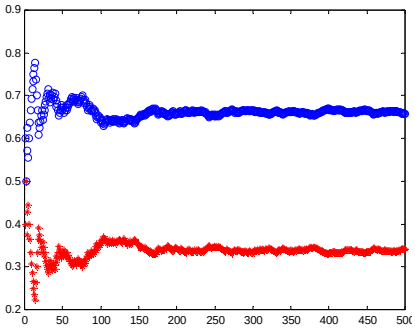


Fig 1(b)

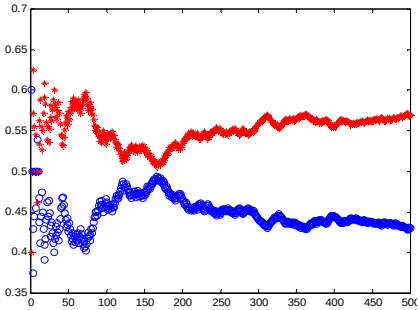


Fig 1(c)

From the result we can see that under the random cases the dissemination of the results is uncertain. The result of shows negative information took the advantage. In Fig 1(b) we can see that negative information took the advantage and positive information still has a certain share after the stabilizing. But the Fig 1(c) shows that positive information took the advantage, even if the negative news accounts on the peak at the beginning. The results well explained the uncertainty of crisis information dissemination.

From the result of Fig 1(a-c), we found that for the study on a random situation is far from enough to draw conclusions. Whether the uncertainty of crisis information dissemination was completely irregular? Of course it's no. In the random process, suppose the number of times to take the ball tends to infinitely, the proportion of red balls and black balls in the jar would tend to a stable value. This stable value is subject to a certain degree of probability distribution. By studying the limit of probability of transmission we can analyze the results of dissemination. Here we defined $p_\infty = p_{1000}$, and took the experiment 500 times, Observe the distribution of p_{1000} in (0, 1). At the first time we got a=20, b=30, c=20, d=10, the result as fig 2 have showed. We got a=20, b=30, c=10, d=10 at the second time, the result as fig 3 have showed.

We can found from the result that if a b c and d had little difference, the impact to balanced outcome by Velocity was clearer than by Initial values. From fig 2 we can conclude that Velocity played an important role to the result. So we can control the disseminate speed of

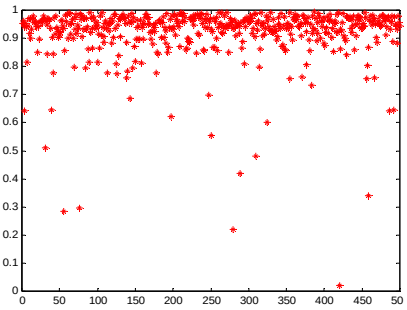


Fig 2 $a = 20$, $b = 30$, $c = 20$, $d = 10$

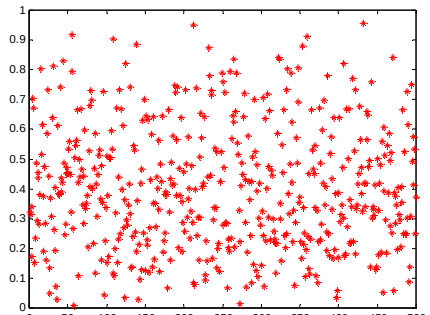


Fig 3 $a = 20$, $b = 30$, $c = d = 10$

negative information to control disseminate results.

C. Matthew mathematical analysis

The crisis information dissemination process is divided into three phases in the above analysis. In the stable period, small adjustments of propagation velocity caused substantial impact of the results of the dissemination. At this time the crisis information dissemination produced Matthew, the re-transmission probability of the one who have take the advantage of the

party was increase. Take ε as the number of red ball after n times test.

$$p(\varepsilon = k) = C_n^k \frac{a(a+c)\cdots[a+(k-1)\cdot c]b(b+c)\cdots[b+(n-k-1)\cdot c]}{(a+b)(a+b+c)\cdots[a+b+(n-1)c]}$$

When n>0 $p(\varepsilon = k)$ subjected to negative binomial distribution.

It represent the probability of failure is x before they succeeded k times when the probability of success is p. This distribution shows that red ball and black ball every time there will be further increase the probability. A successful outcome increased further the chances of success.

Through simulation we can found after a certain number of experiment probability of red balls and black balls had a small magnitude of changes in the propagation velocity c remains unchanged. We call this phase of the crisis information dissemination stabilization period. In the initial stages of the dissemination, no matter red ball or a black ball has a greater impact to the proportion. We call this phase as the sensitive period of crisis information dissemination. The dissemination of crisis information from the sensitive period to the stabilization period, Positive and negative information has gone through a process of accumulation and during the process there was a development stage. Therefore, the dissemination of crisis information can be divided into three phases: sensitive period, development period, a stable period.

III. MODEL INSPIRATIONS

Through simulation we found that the crisis information dissemination has uncertainty. But the crisis communication process can be roughly divided into three stages. Crisis management can effectively avoid the negative information appears on the Matthew Effect at the sensitive period. The model has the following guiding significance for the actual control of the crisis information dissemination.

1). Enhance crisis spreading awareness of the government to eliminate the imbalance in the amount of positive and negative information at the initial period. At the initial period of crisis communication, if the network appears only negative news and government has been slow to intervene in, it would increase the possibility of rumors circulated. Thus, at the initial period of crisis communication, once negative information appears on the network the Government needs to respond in time.

2). Develop rational internet users and control the speed of crisis information dissemination. The disseminate speed of positive and negative information direct impact on the disseminate results. Rational degree of users played a decisive role in the speed of information dissemination. Government needs to carry out crisis education for Internet users in peacetime, develop rational Internet users to distinguish between positive and negative information correctly. In order to effectively control the speed of positive information dissemination.

IV. CONCLUSIONS

This paper built a web-based crisis information dissemination model from the view of Information accumulation, explained the mechanism of dissemination of crisis information and provided with basis for decision-making of crisis management. However, this assumption has controlled many variables involved in the dissemination of crisis information to build a simpler mold. The dissemination model which would consider more factors remains further study to improve the accuracy of the forecasts.

ACKNOWLEDGMENT

This work was supported in part by a grant from the humanities and social science research projects of Guangdong Education Department (its number is 306N5040060) and the Guangdong Philosophy and Social Sciences Planning Projects (its number is B16N4070500).

REFERENCE

[1] Guo Huiming. Crisis communication vs Risk Management [J]. PR Magazine. 2009,(2):92

[2] Garnett, J.A. Communicating throughout Katrina: Competing and Complementary Conceptual Lens on crisis communication[J]. Public Administration Review. 2007, 67, 171

[3] Normal L. Nielson, Anne E. Kleffner, Ryan B. Lee. The evolution of the role of risk communication in effective risk management[J]. Risk Management and Insurance Review, 2005, 8(2): 279-289

[4] Li Zhihong, He Jile, Wu Pengfei. The Time Period Characteristic of Information Communication Model and It's Management Strategies of Paroxysmal Public Crisis [J]. Library and Information Service. 2007, 51(10) : 88-91

[5] Yuan Zhangqiong. The main subjects research on crisis communication care about[J]. Southeast Communication. 2007, 11: 67-69

[6] Perry,D.C, Taylor, M&Doerfel, M.L. internet-based communication in crisis management[J]. Management Communication Quarterly, 2003, 17(2):206

[7] Ren Yuanyuan. The Reestablishment and A pplication of the Modes of Crisis Communication in the Network Era[D]. Shangdong University Master's Thesis. 2008, 4

[8] Fjeld, K&Molesworth. M. PR practitioners' experiences of, and attitudes towards, the internet's contribution to external crisis communication. Corporate Communication; An International Journal, 2006, 11(4): 391-405

[9] Tang Sihui, Yang Jianmei. Study of two diffusion model of mutual information spreading networks[J]. studiees in science of science. 2008, 26(3): 476-480

[10] LIU Chang-yu, etc. Public Opinion Propagation Model Based on Small World Networks[J]. Acta Simulata Systematica Sinica. 18(12), 2006.12: 3608-3610

[11] YU Yong-yang etc. A Study on Public Opinion Evolutionary Model Based on Agent[J]. computer simulation. 2008, 25(9): 9-14.

[12] Liu Jinling. The Genesises of Obstacle about Information Resources Sharing Based on the Information Asymmetry[J]. Modern Information. 2008, 9: 45-47

[13] Ma Feicheng. Etc. Information managenment [M]. Wuhan University Press. 2004. pp75

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 510-513

An Algorithm For NGN Feature Interaction Dectection

Yiqin Lu[1], Guangxue Yue[2], and Jiajin Wang[1]
[1] School of Electronic and Information Engineering , South China University of Technology, Guangzhou, China
eeyqlu@scut.edu.cn
[2] Jiaxing University, Jiaxing, China,
guangxueyue@163.com

***Abstract*—With service design open to third-party in the evolving telecom networks, dynamic method is suitable for feature interaction (FI) detection. When implementing this method, however, the violation of user intension is difficult to express as an indication of FI. To overcome this, a duplicate of Application Server called FRS (Feature-Representation Server) is proposed to add into the network, where each feature can run again individually and the running results are sent back as desirable results to compare with those obtained in the working environment. Interaction exists if they are inconsistent. An algorithm for detecting such kind of FI is given with an experiment for demonstration.**

***Index Terms*—fature ineraction, NGN, algorithm, detecting**

I. INTRODUCTION

The emerging telecommunication networks are service-driven networks, i.e., features will be designed and integrated into the networks in an easier way, and number of service will increase very fast. Here 'feature' is the smallest part of a service provided by the system [1], such as *Call Waiting (CW)*, *Call Forwarding* (*CF*), *Number Portability* (*NP*) , *Originating Call Screening* (*OCS*) etc., and FI is such a phenomenon that two or more activated features interfere with one another and cause abnormality of the system [1-7]. For example, a restaurant with several branches subscribes the feature '*One Number (ONE)*' for delivery service. The feature may connect an incoming call to the nearest branch according to the caller's number. Suppose a man living in *district A* calls the number of delivery service, his call is connected to branch N_A of the restaurant. One day, he removes to *district B*, where the nearest branch is N_B , but he subscribes the feature *NP* to reserve his telephone number used in *district A.* Now, if he calls the number of delivery service, his call is still connected to branch N_A instead of N_B.

Another example of *FI* is the interaction between *CF* and *OCS*. *CF* forwards all incoming calls to another phone. *OCS* prohibits calling to certain callees. When *CF* and *OCS* are both activated in the network, FI may occur as follows: Consider three phones d_1, d_2 and d_3. Suppose d_1 has activated feature *OCS* forbidding a call to d_2, and d_3 has activated feature *CF* to d_2. When d_1 calls d_3, the call is forwarded to d_2, so that d_1 can still connect to d_2, which *OCS* does not expect.

FI existed when the telecommunication system was in PSTN (public switch telephone system) age, and it become more serious when the telecom system evolves to a service-driven network such as IN (intelligent network) and NGN (next generation network), because the number of features increases very fast in a service-driven network.

FI has been studied under three classes of problems: detection [8], avoidance [9] and resolution [10]. Among these classes, detection is investigated the most, because it is the inevitable step in the management of FIs.

FI detection can be carried through in an on-line or off-line manner [3,7,11]. On-line detection is also called static detection. It often uses formal methods such as SDL, FSM, LOTOS, Petri nets [3,12-16] to specify features and verify the system correctness with some criteria. Off-line detection is called dynamic detection. It is realized by capturing the run-time behaviors of the features and comparing with the stored 'correct' behaviors, which could be learned in a test environment [3,7,17-19].

Since the static detection of FI is easier to realize and can find the problem in an earlier stage of a feature lifecycle, most existing literatures are based on this manner [6,12-16]. However, with the telecom networks evolving into NGN, static detection is faced with challenge: in a NGN service-supporting environment, feature design is open to the third-party developers. One can develop a feature using the programming language (like C++ or Java) without getting into the details of network and signaling. For business purpose, the feature logic will be kept private. This makes it hard to specify a feature with formal languages. Therefore, dynamic detection is preferred to use to detect FI in NGN [4, 5,7].

Roughly, the FI can be divided into two kinds [17]:

- Technical interaction: two features run simultaneously and lead the system into inconsistent states or to exhibit inconsistent observable actions.
- Violation of user intention: one or two features fail to meet the user requirements when they run concurrently.

This work was supported by Guangdong Foundation of Science and Technology Project (2006A10101003 2006A1020300, 2008B090500073), Guangzhou Foundation of Science and Technology Project (2003B11609), Project of Technology Breakthrough in Key Fields of Guangdong and Hong Kong (2006Z1) , Natural Science Foundation ofZhejiang province under grant Y1080901

AP-PROC-CS-09CN005

As far as dynamic detection is concerned, the 'technical interaction' occurs when one feature disarrange the procedure of another feature and causes it execute aberrantly, and the 'violation of user intention' occurs when one feature modifies the data or parameter values of another feature and cause it work without desirable results.

For the technical interaction, detection may not be such an arduous task, because there are some structural characteristics in the run-time message sequences when interaction occurs, such as loop, abnormal terminal etc. However, for the intention-violation FI, detection is difficult because it is hard to express the 'undesirable results' with the trivial knowledge learned form individual execution instances of a feature.

This paper proposes a dynamic method for detecting intention-violation FI in a NGN environment. In this method, a redundant application services (AS), named Feature-Representing Server (FRS) in the paper, is added into the network. When a feature is created in AS, a copy of the feature is made in FRS. When two (or more) features are activated together, the trigger conditions of each feature are sent to the FRS. Thus each feature can run individually in the FRS and the result are sent back as desirable result to compare with those result got in the working environment. FI occurs when the tow results are inconsistent.

In the rest of this paper, Section II introduces the service creation in NGN and the challenge of FI detection in it. Section III describes the method: firstly, the additional functional entities of FRS and FIM are introduced, then the format of *SS-FIM message* is described, finally the algorithm of FI detection is presented. Section IV shows the result of implementation of the algorithm. Section V draws a conclusion.

II. SERVICE CREATION IN IN NGN AND THE CHALLENGE OF FI DETECTION

The NGN network is based on a 4-layer architecture, namely Access, Transport, Control, Service. In access layer, the existing telecom networks or user terminals can be connected to a NGN through every kinds of gateway, such as signaling gateway (SG), trunk gateway (TG), network access server (NAS), and access gateway (AG), etc. The core transport network is an IP network. The control of a call or a session is managed by softswitches (SS) in the control layer. The features are running in application servers (AS) in the service layer. Besides the AS, there may be many other servers in the service layer, such as AAA server, network management server (NMS), etc. In such architecture, transport, control and service are separated from each other. That makes the service can be designed by the third-party, who may not know much about the technical details of a telecomm network. The AS may provide an API interface to the service designers so that they can simply design a service with ordinary programming language, such as Java, C++, XML, etc. When a service is designed and created, the feature logic will be kept private for business purpose, which makes it hard to specify a feature with formal languages. Therefore, the traditional static FI detection is not suitable for such a framework. In Section III, a dynamic method for FI detection is present, which is focus on the FI cause by 'user intention violation' (the second type of FI stated in Section I).

III. INTENTION-VIOLATION FI DETECTION BASED ON FEATURE REPRESENTATION

In this section, a dynamic method for detection of intention-violation FI is presented. As shown in Fig. 1, tow entities, i.e. FIM (feature interaction manager) [18] and FRS (feature-representing server), are added into a NGN, and a FIM interface is embedded into SS.

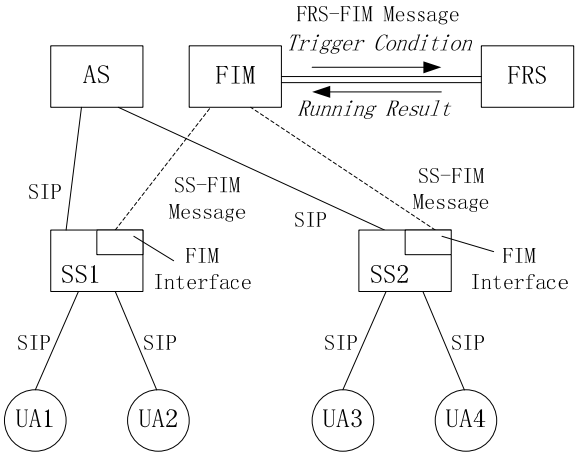


Figure 1. Feature interaction manger (FIM) and feature-representing server (FRS) in NGN

A. FIM and FRS

FIM communicates with SS through 'FIM interface', which captures the SIP flows between ASs and user agents (UA), translates them into so-called 'SS-FIM message' and send it to FIM. FRS is a special AS, which keeps a copy of program of every service when it is created in AS. When two (or more) features are activated together, the trigger conditions of each feature are sent to the FRS. Each feature runs in the FRS individually and the running result is sent back to FIM. FIM uses the result as desirable result of the feature and compare with those getting in the working environment. There may be a FI if the tow results are inconsistent.

There may be more than one AS but only one FRS in the network. In such case, every feature in different ASs must create its copy in the unique FRS.

B. SS-FIM message

The format of a SS-FIM message likes a simplified version of SIP. Only the key fields relating to FI are remained. They are Start-Line, Via Header Field, From Header Field, To Header Field, and Call-ID Header Field. Here, Via Header Field plays an important role. For SS-FIM messages flow from SS to FIM is the hybrid of the signaling of all running features. The Branch ID parameter in the Via Header Field values serves as a

transaction identifier to distinguish which feature the message belongs to.

A call is identified uniquely by *'Call-ID'*. Besides, the FRS-FIM message from FRS to FIM has an additional field representing the results of the feature running individually in FRS.

C. Algorithm for FI Detection

In this sub-section, an algorithm is proposed for detecting intention-violation interaction between two features in NGN. Obviously, how to judge the violation of user intention is the first issue. The algorithm addresses this issue as followings (Fig. 2): once a feature is created in AS, a copy will be sent to FRS. When a feature is triggered in SS, the trigger condition is sent to FRS so that the feature is activated in FRS. The feature runs independently in FRS and the running result is sent back to FIM as the desirable result. At the same time, the feature runs in AS and the actual results can be compare with those sent from FRS. If they are inconsistent, the intention is violated.

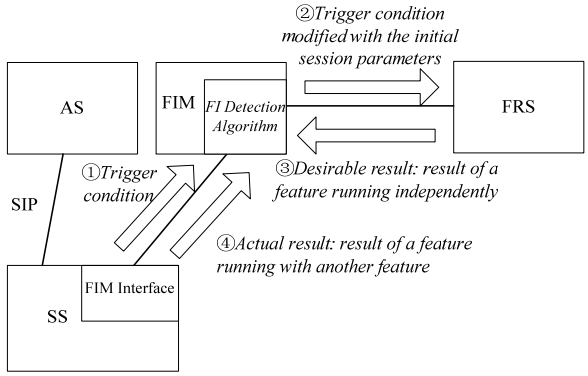


Figure 2. Juage the violation of user intention.

According the intention of which feature is violated, FI may occur in the following tow cases:

1） The intention of second feature is violated, such as the interaction between *NP* and *ONE*. This kind of FI occurs when the first feature changes the initial parameters of the second feature and lead to an undesirable result.

2） The intention of the first feature is violated, such as the interaction between *FW* and *OCS*. This kind of FI occurs when the second feature changes the results of the first feature.

The algorithm runs in FIM. FIM first check if the FI occurs in first case, then check if it occurs in the second case, by executing the following steps:

Step 1. Receive and store FIM-SS messages. If the trigger condition of the second feature arrives before end of the session, go to next step; otherwise exit (, there is only one feature running);

Step 2. Modify the trigger condition of the second feature with the initial session parameters and send to FRS;

Step 3. Store FIM-SS and FIM-FRS messages simultaneously until the desirable result of the second feature running in FRS and its actual result running with the first feature in AS are both received;

Step 4. Compare the two results obtained in Step 3. If they are different, show that the intention of the second feature is violated and FI occurred.

Step 5. Draw the trigger conditions of the first feature from initial parameters and send to FRS.

Step 6. Wait for the desirable result of first feature running in FRS. Compare with the actual result getting in Step 3. If they are different, show that the intention of the first feature is violated and FI occurred.

IV. EXPERIMENT

Experiments have been done by emulating the logical entities in Fig.2, such as AS, SS, FRS, UA, or FIM, etc., each with a PC in an IP network environment. The open-source SIP protocol stack 'oSIP' is employed， which provides SIP parser, URL parser, SDP parser, finite state machine (FSM) and some tools. Above oSIP protocol

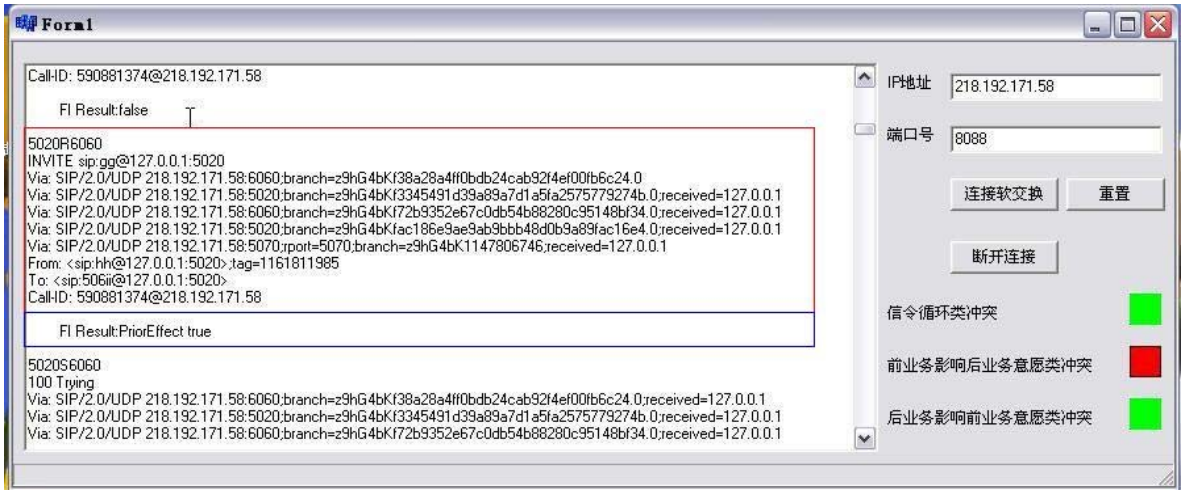


Figure 3. Detecting the intention between CF and OCS.

stack, an open-source network server 'Partysip' is used in the logical entities. Features like *ONE*, *NP*, *UPT* (universal personal telecommunication), *CF*, *OCS*, *TCS* (terminating call screening), etc., have been integrated into AS and FRS.

Fig.3 shows the result of detecting the interaction of *CF* and *OCS*. Assume a user d_1 with SIP URL of 'sip:a@127.0.0.1' activates the feature OCS, forbidding a call to d_2 with SIP URL of 'sip:aa@127.0.0.1:5020', and user d_3 activates feature *CF*, forwarding all the call to d_2. When d_1 calls d_3, it forwards the call to d_3, so d_1 can still make a call to d_2.

In this case, the intention of *OCS* is violated. The FI belongs to second case as is stated above, and is found in the *Step 6* of the algorithm.

V. CONCLUSION

In this paper, a method for FI detection is proposed. The method detects FI during the feature work, and needn't know the feature logic in advance. Thus it is suitable to apply to NGNs. The method focus' on the intention-violation FI and has been implemented in a simulating NGN networks.

Though the method is practicable, it has two disadvantages: one is the increase of cost due to adding of FRS; another is that, if FRS must communicate with SS when running a feature like *UTP*, SS must select a corresponding SS-FIM message stored during the feature running in AS, then if necessary, modify it according to initial session parameters before send it to FRS. The algorithm will be more complex in such case.

REFERENCES

[1] T. F. Bowen, F. S. Dworack, C. H. Chow, N. Griffeth, G. E. Herman and Y. J. Lin, "The feature interaction problem in telecommunications systems", in Proc. 7th Int. Conf. Software Engineering for Telecommunication Switching Systems, pp.59-62, July 1989.

[2] E.J. Cameron, N.D. Griffeth, Y.J. Lin, M.E. Nilson, W.K. Schnure, and H. Velthuijsen, "A feature-interaction benchmark for IN and beyond", IEEE. Commun. Mag., vol.31, pp.64-9, Mar. 1993.

[3] D.O. Keck and P.J. Kuehn, "The feature and service interaction problem in telecommunications systems: a survey", IEEE Trans. Softw. Eng.,vol.24, pp.779-796, Oct. 1998.

[4] W. Bouma and H. Velthuijsen (eds.), Feature Interactions in Telecommunications Systems, IOS Press, 1994.

[5] P. Dini, R.Boutaba and L.Logrippo (eds), Feature Interactions in Telecommunication networks IV. 1997. IOS Press.

[6] Y.Lu, G.Wei, "Feature interaction in telecommunication system", Telecommunication Science, vol.12, no. 12, Dec.1996, pp:8-10

[7] J. Xu, F. Yang, "Feature interaction: present statsu and trends", Telecommunication Science, vol.18, no.9, Sep.2002.

[8] M. Faci and L. Logrippo, "Specifying features and analysing their interactions in a LOTOS environment", In W. Bouma and H. Velthuijsen (eds.), Feature Interactions in Telecommunications Systems, IOS Press, pp.136-151.

[9] M. Jackson and P. Zave, "Distributed feature composition: a virtual architecture for telecommunications services", to appear in IEEE trans Software Engineering.

[10] N.D. Griffeth and H. Velthuijsen, "The negotiating agents approach to runtime feature interaction resolution", in [4], pp.217-235.

[11] S. Reiff-Marganiec and K. J. Turner, "Feature interaction in policies", Computer Networks, vol. 45, pp.569-584, Mar. 2004.

[12] Y. Lu and T.Y. Cheung, "Feature interactions of the livelock type in IN: a detailed example", Proc. 7th IEEE Intelligent Network Workshop, Bordeaux, France, May 1998, pp.175-184.

[13] T. Y. Cheung and Y. Lu, "Detecting and resolving the interaction between telephone features terminating call screening and call forwarding by colored Petri nets", Proc. 1995 IEEE Int. Conf. Systems, Man and Cybernetics, Vancouver, October 1995, pp.2245-2250.

[14] Y. Lu, G. Wei, and T. Y. Cheung, "Managing feature interactions in telecommunications systems by temporal colored Petri nets", in Proc. ICECCS 2001, Skovde, pp.260-270, June, 2001.

[15] I. Zibman, C. Woolf, P. O 拎 eilly, L. Strickland, D. Willis and J. Visser, "An architectural approach to minimizing feature interactions in telecommunications', IEEE/ACM trans. Networking, vol. 4, no.4, August 1996, pp. 582-596.

[16] K. E. Cheng, "Towards a formal model for incremental service specification and interaction management support", In [4], pp.152-166.

[17] S. Reiff-Marganiec, "Runtime resolution of feature interactions in evolving telecommunications systems", Ph.D Dissertation, University of Glasgow, 2002.

[18] S. Tsang and E.H. Magill, "Learning to detect and avoid run-time feature interactions in intelligent networks", IEEE Trans. Softw. Eng., vol.24, pp.818-830, Oct. 1998.

[19] C.Capellmann and K. Kimber, "Toward efficient feature interaction handling", Record of 1996 IEEE Intelligent Workshop, Melbourne, Apr 1996.

[20] S. Tsang and E.H.Magill, "Behaviour based rum-time feature interaction detection and resolution approaches for intelligent networks", in [5], pp.254-270

[21] IETF RFC 3261, SIP: Session initiation protocol. J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler. 2002: 8, 17~18, 26~29, 122

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 514-516

A New Approach for the Dominating-Set Problem by DNA-Based Supercomputing

Xu Zhou[1] , GuangXue Yue[1] , ZhiBang Yang[2] , and Kenli Li[2]
[1]College of Mathematics and Information Engineering, JiaXing University, JiaXing, China
Email: Zhouxu2006@126.com
[2]School of Computer and Communications,Hunan University, Changsha, China
Email: { guangxueyue@163.com , yangzhibang2006@126.com }

Abstract—**DNA computing has been applied to many different decision or combinatorial problems when being proved of its feasibility in experimental demonstration. In this paper, for the objective to reduce the DNA volume of the dominating set problem which belongs to the NP-complete problem, the pruning strategy is introduced into the DNA supercomputing and a new DNA algorithm is advanced. The new algorithm consists of a donimating set searcher, a donimating set generator, a parallel searcher and a minimum dominating set searcher. In a computer simulation, the new algorithm is testified to be highly space-efficient and error-tolerant compared to conventional bruteforce searching.**

Index Terms—**DNA-based supercomputing; dominating set problem; pruning strategy; NP- complete problem**

Ⅰ. INTRODUCTION

The power of parallel, high density computation by molecules in solution allows DNA computers to solve hard computational problems such as NP-complete problems in polynomial increasing time, while a conventional Turing machine requires exponentially increasing time[1]. However, most of the current DNA computing strategies are based on enumerating all candidate solutions[2-13]. These algorithms require that the size of the initial data pool increases exponentially with the number of variables in the calculation, so that the capacity of the DNA computer is limited. And what is more, Fu presented the enumeration algorithms made the length may also too long to make the algorithm to be length-efficient [14].

In this paper, we describe a novel algorithm to solve the Dominating-set problem. Since Huiqin's paradigm proposed in 2004 demonstrated the feasibility of applying DNA computer to tackle such an NP-complete problem. Instead of surveying all possible assignment sequences generated in the very beginning, we use the operations of Adleman-Lipton model and the solution space of sticker, then apply the pruning strategy, a new DNA algorithm for dominating-set problem is proposed.

The paper is organized as follows. Section 2 introduces the Chang et al.'s model in detail. Section 3 introduces the DNA algorithm to solve the dominating-set problem for the sticker solution space. In section 4, the experimental results by simulated DNA computing are given. Conclusions and future research work are drawn in Section 5.

Ⅱ. DNA MODEL OF COMPUTATION

Our novel model employs only mature DNA biological operations. We use the model that took biological operations in the Adleman–Lipton model [1] and the solution space of stickers[13,14] in the sticker-based model in our algorithm.

A. Sticker-based solution space

For the objective of representing all the possible dominating set for the dominating set problem, in our algorithm, vertices are represented by their binary representations using stickers. For every vertex, we denoted two symbols represented by 15-base stickers to encode the information into DNA strands:

$$x_i = \begin{cases} 1 & \text{if the vertex } v_i \text{ is in the dominating set} \\ 0 & \text{otherwise} \end{cases} \quad (1 \le i \le n) \quad (1)$$

Ⅲ. THE NOVEL DNA ALGORITHM FOR SOLVING THE DOMINATING SET PROBLEM

A. Dominating Set Problem[19]

Let G be a graph with vertex-set $V(G)$ and edge set $E(G)$. For any vertex $v \in V$, the neighborhood of v is defined by $N(v) = \{u \in V(G): uv \in E(G)\}$.

Mathematically, a dominating set (DS) of a graph $G = (V, E)$ is a subset $S \subseteq V$ such that each vertex in $V \backslash S$ is adjacent to at least one vertex in S. $|S|$ is denoted as the dominating number. The dominating-set problem is to find a minimum size dominating set in G and has been proved to be a NP-complete problem.

Dominating-set's mathematical model can be described as follows (See Equation 1,2).

$$\begin{cases} f = min\sum_{i=1}^{n} x_i \\ x_j \vee [A(i,j) \wedge x_i] = 1, x_i = 1, j \in \{1,2,...,n\} \end{cases} \quad (2)$$

$$A(i,j) = \begin{cases} 1 & \text{if the vertex } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Now, we will introduce three important theorems about Dominating-set problem in the following.

Theorems 1: The vertex v_i whose degree are one is

not dominating vertex and its adjacent vertex v_j is dominating vertice.

Theorems 2: For a graph G with n vertexes, the dominating number $\eta(G) \le \frac{n}{2}$.

Theorems 3: Using notation $\phi(G)$ to denote the max degree of all the vertexes in graph G *and* $\eta(G)$ to denote the dominating number, we can get the formula $\frac{n}{1+\phi(G)} \le \eta(G) \le n - \phi(G)$.

B. The DNA algorithm for Dominating Set Generator

Based on **Theorems 1** and the definition of Dominating set problem, a new algorithm for constructing the solution space of the dominating set problem is proposed.

Procedure Dominating_Set_Generator(T_0, G, n)

<Input>: Test Tube T_0 and The graph G with n vertexes where n is the number of the vertex.
<Output>: Test Tube T_0 which contains the solution space of the Dominating sets

1: **For** *every vertex* v_i *whose degree is one*
2: *Append*(T_0 , $x^0{}_i$).
3: **For** *the vertex* v_j *that is adjacent to* v_i
4: *Append*(T_0 , $x^1{}_j$).
5: **EndFor**
6: $V = V - v_i - v_j$
7: **EndFor**
8: **For** i = 1 **to** n *where n is the number of the vertex in the set V whose solution space has not produced*
9: *Amplify*(T_0, T_1, T_2).
10: *Append*(T_1, $x^0{}_i$).
11: *Append*(T_1, $x^1{}_i$).
12: *Merge*(T_0, T_1, T_2).
13: **For** *each vertex* v_j *adjacent to* v_i *whose solution space has not produced*
14: *Amplify*(T_0 , T_1 , T_2).
15: *Append*(T_1, $x^0{}_j$).
16: *Append*(T_1, $x^1{}_j$).
17: *Merge*(T_0 , T_1 , T_2).
18: **EndFor**
19: Dominateing_Set_Searcher(T_0 , v_i).
20: **EndFor**

From Dominating_Set_Generator(T_0, G, n), it takes (n-c) *amplify* operations, (cn+2(n-c)) *append* operations, (n-c) *merge* operations where c is the number of the vertexes whose degrees are one, n Dominateing_Set_ Searcher(T_0 , v_i) and three test tubes to construct sticker-based solution space.An n-bit binary number corresponds to an array of input. A value sequence for every bit contains 15 bases. Therefore, the length of a DNA strand, encoding a subset, is 15×n bases consisting of the concatenation of one value sequence for each bit.

C. The Construction of a Dominating Set Searcher

Due to the definition of the Dominating Set problem, a Dominating Set Searcher is designed in the following.

Procedure Dominating_Set_Searcher (T_0, v_i)

<Input>: Tube T_0 includes solution space of all the possible dominating sets for the vertex v_i and its adjacent vertexes.
<Output>: The test tube T_0 of the satisfiable solution space for the vertex v_i and its adjacent vertexes

1: *Extract*(T_0, x_i^1, + (T_0, x_i^1) , - (T_0, x_i^1)).
2: T_1 := + (T_0, x_i^1) and T_2 := - (T_0, x_i^1).
3: **For** j = 1 to $|N(v_i)|$ where $|N(v_i)|$ is the number of elements in $N(v_i)$
4: *Extract*(T_2, x_j^1, + (T_2 , $x^1{}_j$) , - (T_2 , $x^1{}_j$)).
5: T_3 := +(T_2 , $x^1{}_j$) **and** T_4 := - (T_2 , $x^1{}_j$).
6: *Merge*(T_0 , T_1 , T_3).
7: **EndFor**
8: *Discard* (T_4).

From Dominating_Set_Searcher (T_0, v_i), it takes n *extract* operations, n-1 *merge* operations, one *discard* operation and five test tubes.

D. The Construction of a Parallel Searcher

In order to remove the DNA strands which are not the dominating set of the graph G, a parallel clique generator is designed.

Procedure Parallel_Searcher (T_0)

<Input>: Tube T_0 includes solution space of DNA sequences to encode all of the possible dominating sets
<Output>: Test tube T_0 showing all the dominating sets

1: **For** i = 1 **to** n
2: *Extract*(T_0, x_i^1, + (T_0, x_i^1) , - (T_0, x_i^1)).
3: T_1 := + (T_0, x_i^1) and T_2 := - (T_0, x_i^1).
4: **For** every vertex v_j that is adjacent to v_i
5: *Extract*(T_2, x_j^1, + (T_2 , $x^1{}_j$) , - (T_2 , $x^1{}_j$)).
6: T_3 := +(T_2, $x^1{}_j$) **and** T_4 := - (T_2, $x^1{}_j$).
7: *Merge*(T_0 , T_1, T_3).
8: **EndFor**
9: *Discard* (T_4).
10: **EndFor**

From Parallel_Searcher(T_0), it takes n^2 *extract* operations, n(n-1) *merge* operations, n *discard* operation and five test tubes.

E. The Construction of a MiniDominating Set Searcher

Procedure MiniDominating_Set_Searcher(T_0)

<Input>: Tube T_0 showing all the dominating sets
<Output>: Tubes T_i ($0 \le i \le n$) representing the dominating set that contains i vertexes

1: **For** i = 0 **to** n−1
2: $k = min\{i, n/2\}$
3: **For** j = k **down to** 0
4: *Extract*(T_j, $y^1{}_{i+1}$, + (T_j, $y^1{}_{i+1}$) , -(T_j, $y^1{}_{i+1}$)).
5: T_1:= + (T_j, $y^1{}_{i+1}$) and T_j :=-(T_j, $y^1{}_{i+1}$).
6: *Merge*(T_{j+1} , T_{j+1} , T_1).
7: **EndFor**
8: **EndFor**
9: **If** (*Detect*($T_{n/2+1}$) = 'yes') **then**
10: *Discard*($T_{n/2+1}$)**.**
11: **EndIf**
12: **For** i = 1 **to** $\frac{n}{1+\phi(G)}$
13: **If**(*Detect*(T_i) = 'yes'**) then**
14: Discard(T_i).
15: **EndIF**
16: **EndFor**

From MiniDominating_Set_Searcher(T_0), it takes $k \times (k+1)/2 = n \times (n+2)/8$ ($k \le n/2$) extract operations,

$k\times(k+1)/2 = n\times(n+2)/8$ merge operations, $\frac{n}{1+\phi(G)}+1$ detect operations, and $n/2+1$ test tubes.

F. An improved DNA algorithm for Dominating Set Problem

The following DNA algorithm is applied to solve the Maximum Clique Problem

Algorithm 5. Dominating_Set (G, n)

<Input>: The graph G with n vertexes where n is the number of the vertex in G
<Output>: The minimum Dominating set of the graph G with n vertexes

```
1: Dominating_Set_Generator(T0, n).
2: Parellel_Searcher(T0)
3: MinDominating_Set_Searcher(T0)
4: For i = 1 to n/2
5:    If ( Detect(Ti)= 'yes' )
6:       Read(Ti).
7:    EndIf
8: EndFor
```

From those steps in Algorithm 1, the improved DNA based algorithm for dominating set problem can be solved.

G. The performance analysis of the proposed DNA algorithm

The following theorems describe time complexity of Algorithm 1, the number of the tube used in Algorithm 1 and the longest library strand in solution space in Algorithm1.

Theorem 1: The Dominating set problem for any undirected n-vertex graph G with m edges can be solved with $O(n^2)$ biological operations, $O(n)$ tubes and the longest library strand, $O(n)$, where n is the number of vertices in G and.

Ⅴ.CONCLUSION REMARKS

For the objective to reach a free stage in using DNA computers just as using classical digital computers, many technical difficulties such as real time updating a solution when the initial condition of a problem changes, finding out the exact answer quickly and efficiently and the size of the initial data pool increases exponentially with the number of variables in the calculation need to be overcome before this becomes real. We expect our study can make a contribution to clarify that DNA-based computing is a technology that worthwhile us seeking.

ACKNOWLEDGEMENT

This research is supported by the key Project of National Natural Science Foundation of China under grant No.60533010, Natural Science Foundation of Zhejiang province under grant Y1090264, Natural Science Foundation of Hunan province under grant 07JJ6109.

REFERENCES

[1] L. Adleman, Molecular computation of solutions to combinatorial problems. *Science*, vol. 266, 1994, pp. 1021–1024
[2] E. Bach, A. Condon, E. Glaser and C. Tanguay, DNA models and algorithms for NP-complete problems. In: *Proceedings of the 11th Annual Conference on Structure in Complexity Theory*, 1996, pp. 290–299
[3] M. R. Garey, D. S. Johnson. Computers and Intractability: A Guide to the Theory of NP-Completeness, Freeman, San Francisco, 1979
[4] Q. Huiqin, L. Mingming, Z. Hong. Solve maximum clique problem by sticker model in DNA computing. *Progress in Nature Science*. vol.14, 2004, pp.1116-1121
[5] W. L. Chang, M. Guo, H. Michael. Fast Parallel Molecular Algorithms for DNA-Based Computation. *IEEE Transactions on Nanobioscience*, vol.4, 2005, pp. 133-163
[6] W. L. Chang, M.Guo. Molecular solutions for the subset-sum problem on DNA-based supercomputing. *BioSystems*, vol. 73, 2004, pp.117–130
[7] E. Horowitz, S. Sahni. Computing partitions with applications to the knapsack problem. *Journal of ACM*, vol. 21, 1974, pp.277-292
[8] Ho. Michael, W.L. Chang, M. Guo. Fast parallel solution for Set-Packing and Clique Problems by DNA-Based Computing. *IEICE Transactions on Information and System*, vol. E87-D(7), 2004, pp.1782– 1788
[9] Ho. Michael. Fast parallel molecular solutions for DNA-based supercomputing: the subset-product problem. BioSystems, vol. 80, 2005, pp. 233-250
[10] K. L. Li, F.J. Yao, J. Xu, Improved Molecular Solutions for the Knapsack Problem on DNA-based Supercomputing. Chinese Journal of Computer Research and Development, vol.44, 2007, pp.1063-1070
[11] B. Fu, R. Beigel, Length bounded molecular computing. BioSystems, vol. 52 , 1999, pp.155–163.
[12] J. Xu, S.P. Li, Y.F. Dong, Sticker DNA computer model-Part I: Theory. *Chinese Science Bulletin*, vol.49, 2004, pp. 205-212
[13] S. D. Lu. The Experimentation of Molecular Biology. Peking Union Medical College Press.1999
[14] K. H Zimmermann, Efficient DNA sticker algorithms for NP-complete graph problems. *Computer Physics Communications* , vol.144, 2002, pp.297–309
[15] D. F. Li, X. R. Li, H.T. Huang, Scalability of the surface-based DNA algorithm for 3-SAT. *BioSystems*, vol.85, 2006, pp.95–99

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 517-520

Mobile Telemedicine System for Medical Self-rescue

Xiaojun Ma[1], Chunshi Wang[2], Weihui Dai[2] , and Guoxi Li[3]
[1]Shanghai TV University, Shanghai, China
Email: maxj@shtvu.edu.cn
[2]School of Management, Fudan University, Shanghai, China
Email: whdai@fudan.edu.cn, 082025038@fudan.edu.cn
[3]School of Software, Fudan University, Shanghai, China
Email: 073053184@ fudan.edu.cn

***Abstract*—In combination with 3G technology, multimedia technology, database technology, telecommunication technology, etc, as well as the knowledge about medical diagnostics and emergency treatment, this paper presents a mobile telemedicine system for self-rescue in medical emergency. This system can connect the doctors and patients who are physically separated, and thus help the patients realize self-rescue through the two-way audio/video communications. It provides a referential solution to the medical emergency in outdoor accidents and some *Paroxysmal* diseases.**

***Index Terms*—telemedicine, medical self-rescue, outdoor accidents, 3G mobile communications**

I. INTRODUCTION

With the development of economy, the population becomes more fluid, and in the meanwhile there is also increasing requirement for medical service at anytime and anywhere, especially in mobile environment.

When an outdoor accident or *Paroxysmal* disease happens, the earliest and most effective rescue is offered not by the medical staffs, but by people themselves. At that moment, if the first-aid instructions can be provided by medical experts, the casualties are expected to be minimized. This led to the development of telemedicine which applied telecommunication technology, computer multimedia technology and other information technology to transmit medical information for diagnosis, monitoring, treatment and education. The medical information usually includes medical images, real-time audio/video, patient records, data outputted by medical equipments, etc.

In 1967, Massachusetts General Hospital (U.S.) first established a telemedicine system. In 1986, Mayo Clinie created the earliest commercialized telemedicine system. In 1991, the Telemedicine Centre of Medical College of Georgia are founded, all rural hospitals and clinics within the state could get access to this centre and got special medical aid from it. In the late 1990s, the forms of telemedicine and mobile medicine diversified, such as mobile home monitoring, mobile electronic records, mobile rescue, etc. The U.S. and Europe were advanced in this field. In their countries, huge investment was laid out on the main direction of tele-consultation and tele-treatment. HIS (Hospital Information System) and EPR (Electronic Patient Record) technology had already been almost perfect in those regions. At the same time, Japan had already constructed large-scale mobile medicine facilities, which covered a great proportion of its land. European Union committed to the establishment of laboratories, and had carried out several large scale experiments on telemedicine. Australia, South Africa, U.K., Singapore and other countries also made great effort in mobile medicine field.

The development of telemedicine and mobile medicine was relative backward in China. We had not set foot in this field until late 1990s. In 1997, China's Golden Health Care Network formally put in to operation to support the health and related ministries. In 2003, the experts group of the Chinese PLA 253 Hospital first used tele-consultation for military expert consultation. With the development of mobile communication, various studies on GPRS-based mobile medicine took place, which are mainly for community mobile medicine and monitoring. These schemes were on the basis of 2.5G. This concept of mobile telemedicine for emergent self-rescue origins from nowadays rapid development of 3G mobile communication technology, networking technology, and multimedia technology, in combination with the previous telemedicine technology. Through effective integration of these technologies, we hope to establish a complete scheme of mobile telemedicine system that can be accessed via 3G network, to help people realize self-rescue while they are far away from the traditional medical service.

This paper proposes a concept of mobile self-rescue for medical emergent. Emergent self-rescue is in demand when accidental injury or *Paroxysmal* disease occurs. Accidental injury means physical damage or hurt caused by different kinds of sudden incidents or accidents, including physical, chemical and biological factors. If the public can get helpful first-aid when accidents happen, the casualties can be minimized. Essentially, the mobile telemedicine service for emergent self-rescue belongs to the category of mobile medicine. Telemedicine is an applied subject, using various technologies to transmit

This research was supported by Shanghai Leading Academic Discipline Project (No.B210).
Corresponding author: Weihui Dai.

AP-PROC-CS-09CN005

medical information for diagnosis, monitoring, treatment and education. The medical information includes medical images, real-time audio/video, patient records, data outputted by medical equipments, etc.

II. SYSTEM DESIGN

A. Syestem structure

This mobile telemedicine system for medical self-rescue is an open distributed system. It is composed of six parts: 3G client (user client/call-response client), Call-response Subsystem, E-expert Subsystem, Geographic Information Subsystem, Database Subsystem, and the Response System in medical institutions, as shown in Figure 1.

These systems are organically combined to fulfill the self-rescue process, which includes rescue initiation, response, solution selection, deployment, GIS positioning, etc.

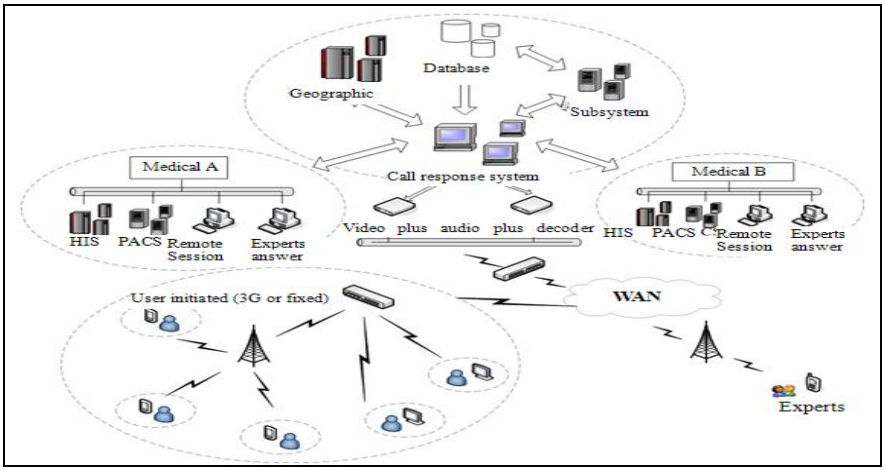


Figure1. Mobile telemedicine system for medical self-rescue

B. Subsystems

Besides 3G clients, which can be realized through secondary development to the module of 3G phone, there are mainly four subsystems in this mobile telemedicine system.

1.Call-response Subsystem. Call-response System is the entrance to the whole system through 3G client. It delivers the required information from E-expert Subsystem or real experts directly to users through computer automatic answer equipments or by the operators, providing salvation instruction and the related information. It plays the role of "information provider" in the whole system, and enables the service demanders to obtain useful information conveniently and quickly. It is a crucial system which connects various subsystems and provides interactive services for users through all kinds of modern communications.

The flow chart of call-response system is shown in Figure 2.

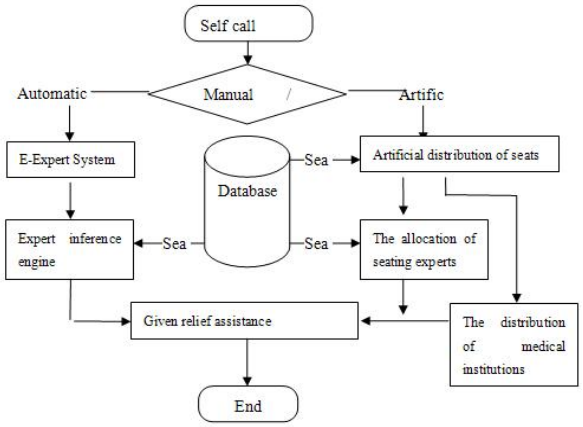


Figure2. Flow chart of the call-response system

2.E-Experts Subsystem. Medical problems are usually complex, and it is not unusual to be trapped in the dilemma while information is uncertain. E-expert system can utilize the relative experience and rules in obtaining reasonable judgments, providing suitable solutions, and making effective forecast. With the 3G mobile communication, expert system can deal with the problems more effectively and quickly.

3.Database Subsystem. The database subsystem provides data support to various subsystems. It provides the seats and online experts' information to call-response system, which makes it possible for the call center to connect to right information source quickly. And it combines with the E-expert subsystem to offer comprehensive data to support the expert network. GIS system is also connected.

4.GIS subsystem. GIS is mainly used to identify the geographic information from the caller end.

III. CORE TECHNIQUES

The design and implementation of this system require the coordination of various kinds of technologies. Here we briefly introduce some core techniques.

A. 3G's interflow with fixed network

3G's interflow with the fixed network is definitely in huge demand as the popularization of handheld terminals and high speed bandwidth. 3G standards set 3G-324M as

the protocol standard for communication systems and terminal devices; MPEG-4 is recommended to be the video protocol standard. While in fixed network, most video business is on the basis of H.323 in packet network and SIP. Protocol H.323 is the Internet and LAN communication standard, made by ITU-T. SIP is the multimedia circular order protocol standard, set by IETF. The interflow process is as follows: 3G end initiates the requirement for connection, and build conversation links with 3G gateway (VIG) through 3G-324M protocol; then VIG build conversation links with IP network via H.323 protocol, or with soft-exchange network via SIP. The transformation of protocols is done in VIG, which perfectly overcomes the bottleneck in implementing the interflow between 3G and fixed network.

B. Streaming Media Technology

In wireless environment, because the influence of multi-path attenuation, noise and other factors, the channel error rate is usually high and the error rate changes with the external environment. Therefore, the video/audio data flow should be compression coding in 3G network transmission process, to ensure data quality and display speed during decoding. The transmission form of video/audio frequency in 3G network is streaming media. It decomposes the source into small packages, use cache technology to weaken the effect of delay and vibration, and ensure the order of these packages.

C.Computer Telecommunication Integration

CTI supports a wide range of operating systems such as WINDOWS/LINUX. It is featured with high reliability, large connect capacity (of single point) and great expandability. CTI products are established on the distributed architecture, and are composed of modules. CTI middle-ware offers a standard platform, supporting ordinary calls, IP calls, Email calls, SMS calls, Web calls and other calling channels. CTI middle-ware is applied to both switch mode call centers (such as Avaya, Alcatel, Nortel, Siemens), and card mode call centers. The whole calling process is continuously managed and monitored.

IV. APPLIED EXAMPLE

The 3G-based mobile medicine system for emergent self-rescue can be applied widely. Here an example of sports injury is provided, and by comparison with the traditional 120 first-aid, the new emergency treatment measure has its unique advantages.

Sports injuries are injuries that occur to athletes participating in sports events. Bruise, muscle strain, ligament sprain, dislocation and facture are common sports injuries. In recent years, with the popularization of sports, the incidence of sports injury increases. The harm of sports injury is nonnegligible. Inappropriate or late treatment will cause great pains to the injurer; what's more, sports injury may deteriorate into permanent or chronic diseases. Therefore, effective and timely treatment to sports injury is very necessary.

A. 120 first-aid system

Traditionally, when sports injury occurs, people first resort to 120 first-aid. However, with the development of modern economy and the rapid progress of urbanization, the increasing density of buildings and traffic make it difficult to carry out timely and effective medical treatment at the accident scene. This is mainly due to inaccurate or late report, traffic block, or inappropriate treatment by medical staffs. Besides, 120 first-aid system itself has many shortages, such as outdated equipments, slow response, lack of geographic positioning abilities and so on. The traditional 120 first-aid process is described in Figure 3.

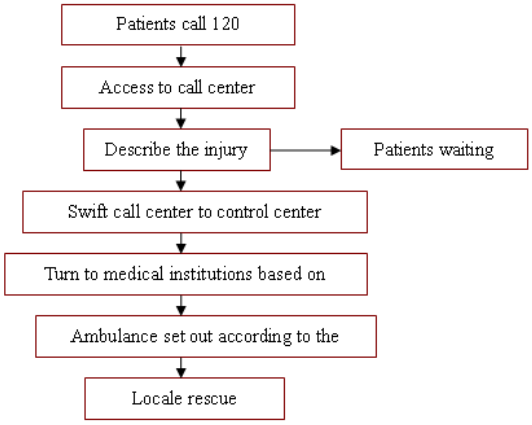


Figure3. The workflow of the traditional 120 first-aid

In the present 120 first-aid system, communication about the injury situation and location is totally based on speech description which may suffers from description error and misunderstandings that lead to false decision. On the other hand, the emergency treatment decisions are made by manpower on the basis of personal experience, no standard information support is provided. And sports injuries often happen in the city periphery, which is difficult for 120 first-aid to arrive the scene timely. So a more effective, timely, and standard first-aid solution is in demand.

B. 3G-based mobile medicine system

3G-based mobile medicine system for emergent self-sescue is greatly different from the traditional 120 first-aid system. It is a comprehensive system composed of 3G communication, video/audio, GIS, call center system and the like, and takes the advantages of 3G technology to guide and supervise the client end.

By using this mobile medicine system, the medical staffs and experts can make quick and accurate situation assessment, for they can not only hear from the injurer, but also see what happens and the surrounding environment. This system also helps to give effective on-site guidance for emergent treatment, thus creates a green life channel for the patients.

The mobile medicine first-aid process is presented in Figure 4. As it shows, all steps are bidirectional, which means continuous communication between the patients and the medical service provider.

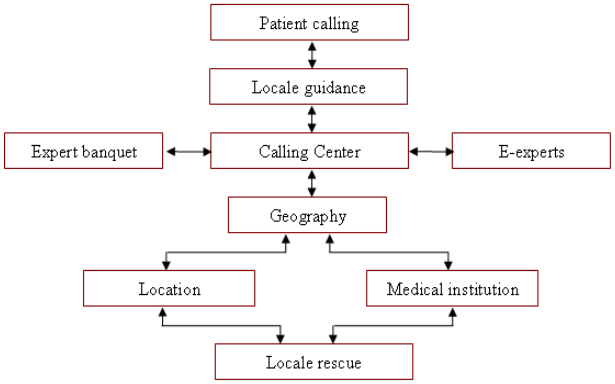


Figure4. The workflow of 3G-based mobile medicine system for emergent self-rescue

The advantages of 3G-based mobile medicine system for emergent self-rescue are: (1)it can check the on-site situation effectively by 3G audio/video methods, and make the most direct diagnosis about the injurer's situation; (2)the switches between operator seats and professional (expert) seats can help offer a more authoritative guidance and advice; (3)it can make optimal path analysis within certain geographic range around the caller's exact position; (4) rescue decisions are more scientific and reliable by real time positioning and navigation, along with affluent information from both online experts and E-experts.

V. CONCLUSION

3G, as the new generation of data communication technology, is and will be applied in various fields, such as communication, medicine, manufacture, living and so on. This paper analyzes and designs a 3G-based mobile medicine system for emergent self-rescue. Also, an applied example is provided, and compared with traditional 120 first-aid method. The theoretical and practical significance of this paper are:

Research on 3G and related technologies in detail, and make perfect combination of these technologies to set up a new 3G system; crucial and difficult techniques are considered, such as streaming media, 3G compatibility problems and E-expert system.

Design a 3G-based mobile medicine system on the basis of 3G communication technology, streaming media technology, neural networking technology, database technology, as well as GIS and CTI.

Explore a new and meaningful application field of 3G technology and mobile medicine. Through this 3G-based mobile medicine system, traditional emergent treatment barriers are smoothed, the bidirectional video/audio communication supported by abundant and accurate information database makes the emergent self-rescue not only possible, but also timely and effective.

Introduce E-expert system to the self-rescue system. The traditional simple "person-person" self-rescue mode is extended by "person-machine" mode, which greatly increases the efficiency and the accuracy. What's more, the BP neural networking enables E-expert system with the ability of self-learning, thus continuously enhances the diagnosis level.

The 3G-based mobile medicine system for emergent self-rescue is a comprehensive application of various technologies in medical self-rescue field. It is an interdisciplinary research area, and related researches are both at in and abroad. This paper is an active exploration and a beneficial attempt to the field. With the popularity of 3G terminals, mobile medicine systems have great market potentials. Still, there are issues calling for in depth study and consideration, such as the information safety problem, the profit model, technology standards uniformity, etc.

ACKNOWLEDGMENT

This research is supported by Shanghai Leading Academic Discipline Project (No.B210).

REFERENCES

[1] Nagy K. "Telemedicine Creeping into Use, Despite Obstacles," *Journal of the National Cancer Institute*, Vol. 86(21), pp.1576-1578, 1994.

[2] Egan GF liu IQ. "Computers and Networks in Medical and Healthcare Systems," *Computers in Biololgy and Medcine*, Vol.25 (3), pp.355-365., 1995.

[3] R.S.H. Istepanian, B. Woodward, and C. I. Richards. "Advances in telemedicine using mobile communications," *IEEE*, Vol.1, pp. 3556-3558, 2001.

[4] Berkman.P, Heinik.S, Rosenthal.M, and Burke.M.. "Supportive Telephone out Reach as an Interventional Strategy for Elderly Patients in a Period of Crisis," *Soc-Work-Health-Care*, Vol.28 (4), pp. 63-76, 1999.

[5] Chen Huiming, and Li Yanhua. "Technology of MCU-controlled GSM Phone and its Application," *Microcontrollers & Embedded Systems*, Vol.2, pp.12-14, 2005.

[6] Louis Lareng. "Telemedicine in European," *Journal of Internal Medicine*, Vol.13, pp.1-3, 2002.

[7] R.S.H. Istepanian, B. Woodward, and C. I. Richards. "Advances in telemedicine using mobile communications," *IEEE*, Vol.1, pp. 3556-3558, 2001.

[8] Xiong Wu, and Ma Binrong. "Analysis of the present situation and the trend of medical diagnosis expert system," *Information of medical Equipment*, Vol.4, pp.27, 2006.

[9] Yang Tao, Zheng Xiaoxia, and Liu Jingde. "Research and Implementation of CTI Middleware Based on CSTA Protocol," *Journal of Computer Applications*, Vol.10, pp.14-16, 2001.

[10] Zhang Hua, Wang Chongjun, Ye Yukun, and Chen Shifu. "SARSES: Design and Implement of an Expert System for Diagnosis," *Computer Engineering and Applications*, Vol.18:217, 2004.

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 521-524

Integration Middleware for Mobile Supply Chain Management

Weidong Zhao[1], Haifeng Wu[1], Weihui Dai[2] , and Xuan Li[1]
[1] School of Software, Fudan Universtiy, Shanghai 200433, P.R.China
Email: {wdzhao, 082053025,082053002}@fudan.edu.cn
[2] School of Management, Fudan University, Shanghai 200433, P.R.China
Email: whdai@fudan.edu.cn

***Abstract*—According to the mobile supply chain management features and considering the limitations and obstacles in the mobile environment. This paper proposes a multi-agent integration middleware for mobile supply chain management, aiming to solve information integration and data synchronization problem without worrying about network or terminal heterogeneity in mobile environment, which can achieve mobile supply chain dynamic integration and then improve the efficiency of supply chain management. Compared with traditional distributed approach, the management method has lots of advantages, such as saving network bandwidth; its flexibility and scalability are also very favorable.**

***Index Terms*—multi-agent middleware, mobile supply chain, semantic heterogeneity, intelligent agent, ontology library**

I. INTRODUCTION

In today's global, competitive and dynamic business environment, the competitions among enterprises have transformed from company versus company to supply chain management versus supply chain management [1]. The objective of a favorable supply chain management should ensure to deliver the right product, at the appropriate time, at the competitive cost, and with customer satisfaction for keeping the competitive advantages [2]. With the development of communication technologies and mobile networks, the supply chain management in the mobile environment has become more and more prevalent and necessary. Nowadays, more and more enterprises have paid or are paying attention to replace traditional linear supply chains with mobile and adaptive supply chain management, helping firms to reduce management cost and gain supply chain responsiveness or other competitive advantage [3]. However, compared with the traditional supply chain management, there are also some limitations existing in the mobile environment. For example, storage ability and processing ability of mobile terminals are so restricted and wireless networks maybe are heterogeneous and changeful, so its management needs a simple and efficient method for data integration and information sharing [3]. Multi-agent technology is considered as a favorable method and has achieved many remarkable results in many domains because of its advanced mentality and great describing ability. The adaptability, self-government and sociality characteristic can help agents work well in changeful, un-structural and complex environment and finally achieve real-time communication and decision establishment [3]. In this paper, based on multi-agent middleware technology, an agile supply chain management method in mobile environment is proposed according to the transforming orientation of supply chain management in manufacturing domain, aiming to achieve data synchronization, message transferring, services management and application integration, which can shield off communication network or mobile terminal heterogeneity and offering lots of flexible interfaces for external application program by integrating many static or mobile agents. The supply chain entities can join the supply chain management activities conveniently with mobile terminals or other mobile communication devices.

The rest of the paper is organized as follows: background reviews, previous works between multi-agent technology and mobile supply chain coordination management are introduced in Section 2. In section 3, we briefly describe the structure of the multi-agent middleware and discuss how to design it. Technology implementation and advantage of multi-agent middleware is analyzed in section 4. Finally, we will conclude this paper with future work.

II. RELATED WORK

With the development of wireless network and communication technology, traditional supply chain has been upgrading to mobile supply chain [4]. Certainly, research on the supply chain management in mobile environment has become very important and significative. How to design a flexible and agile supply chain management method in mobile environment has received considerable attention both in research and in practice [4]. Mobile supply chain management, which was proposed in the end of the last century, is integrated with mobile terminal devices, advanced mobile communication technologies and wireless network technology [5]. The essence of mobile supply chain management is to ensure anyone can take necessary activities or utilize existing services at any time and any place.

Compared with traditional supply chain management, the mobile supply chain management has many obvious advantages, for instance, mobile supply chain

This research was supported by National High-tech R & D Program (863 Program) of China (No.2008AA04Z127) and Shanghai Leading Academic Discipline Project (No.B210).
Corresponding author: Weihui Dai.

AP-PROC-CS-09CN005

management is very appropriate for capturing and managing the continuously changing information or resources in time, it can also reduce the limitations of information interactivity and offer real-time services for the supply chain entities. High automation, simplification, standardization, and modularization also can improve the efficiency of processing critical information and reduce the uncertainty of supply chain management [4].

Undoubtedly, there are also some limitations and obstacles in mobile supply chain, first, the heterogeneity among mobile terminals or wireless network are inevitable. Besides, the limiting bandwidth of mobile networks and the limiting ability of mobile terminals also can affect the efficiency and speed of information transferring. Therefore, how to achieve mobility and overcome the limitations in mobile environment should be considered carefully. Many scholars have found that previous integration technologies, such as EDI, E-Hubs, can not be very helpful for these problems, but multi-agent can be regarded as an efficiency method because of its advanced mentality, autonomy, adaptability, and sociality, which can be designed as a singer layer in the mobile supply chain management system, supporting dynamic interaction for the supply chain entities and creating optimal supply chain organization scheme [6].

III. MULTI-AGENT INTEGRATION MIDDLEWARE FOR MOBILE SUPPLY CHAIN MANAGEMENT

As research shows, agent technology can well carry out the supply chain behaviors by inter-operation across the mobile network nodes at the abstractive level in a computational system because of its great potentials in supporting supply chain management. Multi-agent integration middleware is for managing and integrating many different functions agents uniformly and offering a flexible and favorable communication bridge between mobile terminal and external application. In the multi-agent middleware, every agent takes charge of its own functions and offers its own services for other agents. The whole structure of mobile supply chain can be divided three layers, naming distribution service layer, operation function layer and enterprise resource layer. Every layer is controlled by lots of corresponding agents and the agents in every layer have their own functions. For example, agents in distribution service layer are responsible for receiving order requests from customers, and then distribute special information or services for customers' requests. The Multi-agent integration middleware in our paper, containing all these agents, can integrate the functions of these different agents and furnish fine communication strategies for different agents, offering many kinds of personalized services to external application with uniform interfaces. The function of multi-agent middleware is to create optimize supply chain organization scheme, and then implement personalized and efficient mobile supply chain management. The architecture of the multi-agent middleware is shown in Fig. 1.

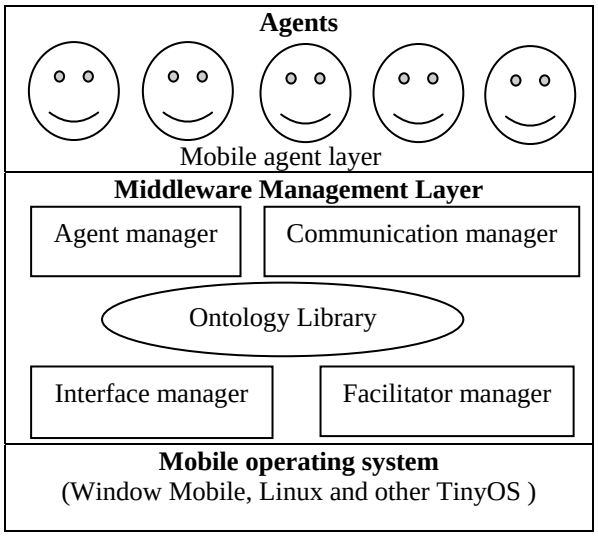


Figure 1. Framework of the multi-agent middleware

The structure of the middleware is also divided into three layers. The top layer is called agent layer, which contains many static or mobile agents, being responsible for creating independent agents in accordance with specific business necessities and offering a comfortable working environment for them. In this layer, every agent can finish its own function independently. For example, order agent can receive the order requests from customers placed via mobile terminal device and then send them to specific merchant agent, waiting for response or processing result. The second layer, as the management layer, contains the core components for managing all agents of the integration middleware, such as agent manager, communication manager, ontology library, interface manager, facilitator manager and so no. The agent manager module is responsible for managing agent registration and allocating a specific amount of memory and unique ID for these agents. In the communication manager module, the communication behavioral rules and interaction strategies between different agents are maintained. The facilitator manager module offers yellow-page services and orientation services for agents in the middleware. For improving the flexibility and expansibility of the multi-agent middleware, many uniform interfaces were designed to integrate with external application, such as web application interface, wireless application interface. While an ontology library is built to solve semantic heterogeneity problem during the process of transferring data and information in mobile network [7]. Operating system layer, as the nearest layer to mobile terminal, it aims to shield off the heterogeneity among different mobile terminal systems or communication networks. Considering the bandwidth limitation in the mobile and wireless network, we divide each agent into tiny packets to transmit and offer retransmitting measure, in order to minimize the bad impact of message loss, which is quite common in mobile and wireless network

In our paper, multi-agent middleware is used to achieve agile mobile supply chain management in manufacturing domain, whose framework is shown in

figure 2, aiming to realize services integration, process arrangement and data synchronization. For the sake of goal, we designed lots of agents; they are divided into two kinds: mobile agents and static service agents, static service agents only execute in the fixed computing nodes, being responsible for dealing with services requests, while mobile agents can move freely across the mobile network and communication with static service agents. Moreover, mobility design and asynchronous data communication mode are implemented with efficient methods.

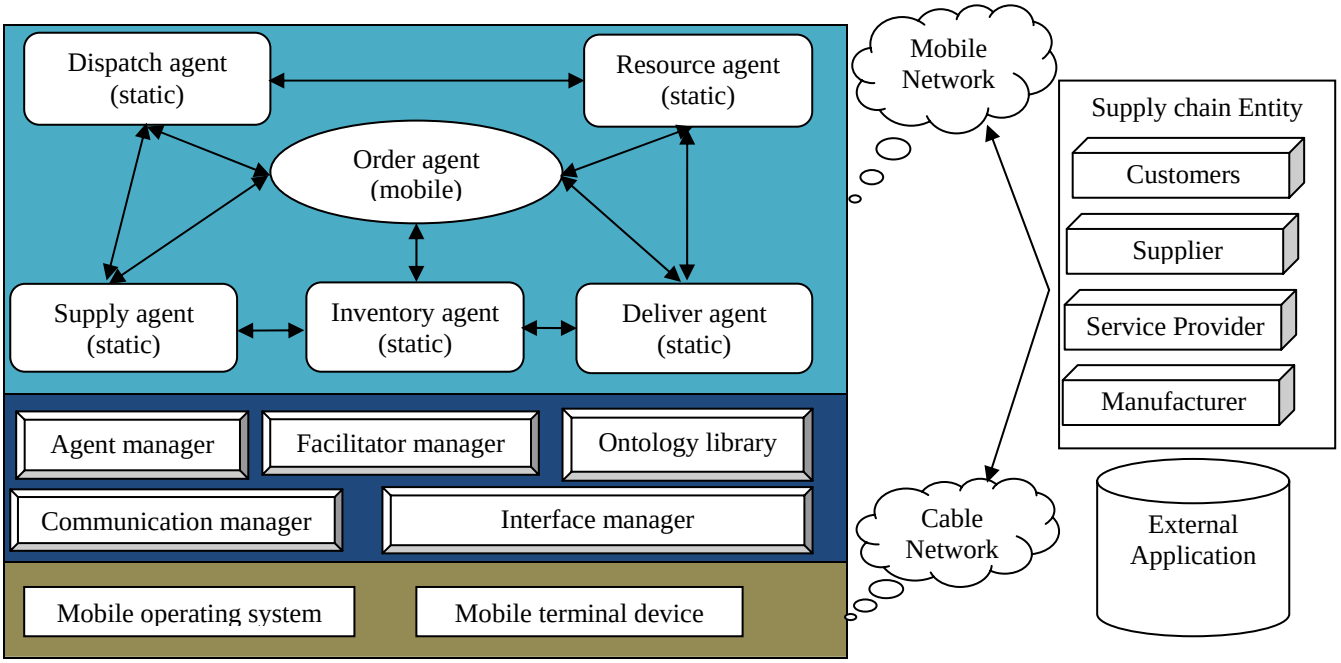


Figure 2. Mobile supply chain management framework based multi-agent integration middleware

In the middleware, dispatch agent, supply agent, resource agent, deliver agent and inventory agent are designed as static services agents. Dispatch agent is responsible for searching coincident supplier according to the information of customer order; supply agent is responsible for describing the resources and services provided by suppliers, resource agent is responsible for describing and managing dynamically the resources of mobile network, deliver agent is responsible for producing a serial of delivering statement, and inventory agent is responsible for checking the inventory status of supplier or manufacturer. At the same time, order agent is designed as a mobile agent, moving between the different network nodes with corresponding request, which is responsible for transferring order request and capture corresponding response and resources.

Well-behaved communications between different agents is another characteristic of the multi-agent middleware, different agents make communication with ACL messages, being integrated the communication content for agents and offer convenient interaction measures. On the basis of FIPA explanation, agents carry through communications by asynchronous message transmitting. Practices have shown that asynchronous message transmitting mode is very useful and efficient in mobile and wireless network, because it allows senders disconnect network temporarily after sending information. For example, when a customer places an order through mobile terminal or other wireless network equipment, order agent will capture the order information and then move in the mobile network, in order to find to requiring agent to transfer information. With the help of interactions between different agents, an optimized mobile supply chain organization scheme will be created and customers can obtain personalized services from the management system. By using real-time information sharing in agent management module, the design method cannot only prevent the Bullwhip Effect in traditional supply chain management, but it also can improve the efficiency of mobile supply chain management [8].

V. MIDDLEWARE TECHNOLOGY IMPLEMENTATION

In order to actualize the multi-agent middleware and then apply it in mobile supply chain management well, we deploy the multi-agent middleware in a fixed network, providing middleware services for mobile terminal and offer parsing functions for upper level application. After comparing many typical multi-agent system development platform, we choose JADE (Java Agent Development Framework), to develop multi-agent in this middleware, with its integrating communication strategy, interaction protocols, and other these characteristics in this development platform, we achieve integration and communication among agents well without worrying about the troubles brought by systems or terminals heterogeneity. First, we design a lot of agents, including order agent, dispatch agent, supply agent, resource agent, deliver agent, inventory agent and so on. Only the order

agent is mobile agent, moreover. Every agent obtains one valid AID (directory of agent identifiers) after registering with agent management module, and agent management module offers yellow-page services and lifecycle management for these agents. Mobility and communication strategy also are well achieved with measures and protocols integrated in the platform. In order to implement mobile terminal user management, we design a user profile in middleware to store user information, including much basic registering information and some personalized information. By conserving agent executing status and data information, discovery service and location management are finished. As for resource management, we assign resource management object for every agent, maintaining and managing data, threads and so on. Besides, standard security interface is used to support data encryption, digital signature and other protecting measures.

For achieving uniform description for resources, we choose Resource Description Framework to dispose data with different structures. First, we execute meta-data abstraction, and then deal with these meta-data to describe the resources of different network nodes as a flexible but uniform style. On the mobile terminal, we design and implement favorable interfaces for mobile applications, users can join in the mobile supply chain management activities by using legitimate account, and then the corresponding agent will move in the mobile network to communication with other agents for returning responses for customers after the multi-agent middleware execute mobile supply chain coordination management activities.

In order to validate the practicability and reliability of this multi-agent middleware, we apply the management method in a small case study. In this case, customers submit order for buying flowers with their concrete requirement, including flower name, desirable price and other information. After the order information is submitted, order agent starts and moves into the mobile network zone covered with this middleware to find service suppliers for its request. The order agent will make communication with other agents according the request information. Finally, the multi-agent middleware will return the response to terminal user. Generally, mobile terminal users can receive efficient and customized services. Through this simulation experiment, we can find the supply chain management method based on multi-agent middleware can really improve the efficiency for enterprise management and customers can obtain more satisfied services by using the good features of agent and reasonable communication strategies. Furthermore, the troubles existing in the traditional supply chain management, such as bullwhip effect, also have relieved.

VI. CONCLUSION

This paper presents a mobile agent middleware specifically designed for agile mobile supply chain management. After considering the characteristic and the limitation of the mobile and wireless network and comparing mobile supply chain management with traditional supply chain management, we give a brief design of a multi-agent middleware, analyzing its framework and discuss how to implement it using many existing technology. This mobile supply chain management model really can improve flexibility and reaction speed of the whole supply chain. Furthermore, it can solve integration problem between different mobile equipments and distributed storage, which can offer a new technology measure for mobile integration problem. Compared to previous distributed technology, this method has its own advantage and the further research is to integrate this multi-agent middleware based mobile terminal with the existing application program well

ACKNOLEDGMENT

This research is supported by National High-tech R & D Program (863 Program) of China (No.2008AA04Z127) and Shanghai Leading Academic Discipline Project (No.B210).

REFERENCES

[1] Barbuceanu M., and M.S. Fox. “Coordinating multiple agents in the supply chain,” In Proceedings of the Fifth Workshops on Enabling Technology for Collaborative Enterprises, pp,.134-141.1996.

[2] Yonghui Fu, and Rajesh Piplani.,“Multi-agent enabled modeling and simulation towards collaborative inventory management supply chains,” Proceedings of the 2000 Winter Simulation Conference, pp. 1763-1771, 2000.

[3] Ercan Oztemel, and Esra Kurt Tekez, “Interactions of agents in performance based supply chain management,” Journal of Intelligent Manufacturing, Volume. 20, pp.159-167, 2009.

[4] Qi Yuan, Zhao Xiaokang, and Zhang Qiong. “Key Technology and System Design in Mobile Supply Chain Management,”International Symposium on Electronic Commerce and Security, pp.258-264,2008.

[5] ChangYang and Yu-qiang Feng, “Integrated Multi-Agent-Based System for Agile Supply Chain Management,” Machine Learning and Cybernetics, pp. 23-27,2006.

[6] Milind Tarnbe, J. Santofimia, Francisco Moya, and Felix J. Villanuev, “Integration of Intelligent Agents Supporting Automatic Service Composition in Ambient Intelligence,” Web Intelligence and Intelligent Agent Technology, pp.504-507, 2008.

[7] Chien-Liang Fok, Gruia-Catalin Roman, and Chenyang Lu, “Agilla: A mobile agent middleware for self-adaptive wireless sensor networks.,” ACM Transactions on Autonomous and Adaptive Systems, pp.1-26, 2009.

[8] Siddharth Mundle, Nupur Giri, and Arpita Ray, Shrikant Bodhe, “JADE based Multi Agent system for mobile computing for cellular networks,” Proceedings of the International Conference on Advances in Computing, Communication and Control,pp.467-473, 2009

ISBN 978-952-5726-07-7 (Print), 978-952-5726-08-4 (CD-ROM)
Proceedings of the Second Symposium International Computer Science and Computational Technology(ISCSCT '09)
Huangshan, P. R. China, 26-28,Dec. 2009, pp. 525-528

Mobile Agent System for Supply Chain Management

Wenjuan Wang[1], Tong Li[1] , Weidong Zhao[2], and Weihui Dai[3],
[1] School of Software, Yunnan University, Kunming 650091, P.R.China
Email: wenjuan42@gmail.com, tli@ynu.edu.cn
[2] School of Software, Fudan University, Shanghai 200433, P.R.China
Email: wdzhao@fudan.edu.cn
[3] School of Management, Fudan University, Shanghai 200433, P.R.China
Email: whdai@fudan.edu.cn

***Abstract*—In the 21st century, competition between companies has been turned into the competition between their supply chains. To satisfy the rapidly changing demands in global market, agile capability and flexibility have been the new requirements in today's Supply Chain Management (SCM). Some intelligent activities, such as negotiation, decision and collaboration, have become the "bottleneck" to improve the performance of SCM. This paper first discusses the development of mobile agent and explores its application in SCM, and then presents a mobile agent system for SCM in the case of flower trading. It can deal with the negotiation, decision and collaboration intelligently and automatically.**

I. INTRODUCTION

The primary objective of Supply Chain Management (SCM) is to fulfill customers' demands through the most efficient use of resources, including distribution capacity, inventory and labor [1]. Nowadays, customers become more demanding in competitive and global market. A single enterprise is no longer capable to deal with the fast changing and customized demands, so the SCM-based alliance is becoming the mainstream of business organization in the 21st century. Therefore competition in today's market is no longer of company versus company but rather the supply chain versus supply chain [2]. The performance of SCM has become the decisive factor to satisfy various demands with the lowest cost.

A supply chain is a system of organizations, people, technology, activities, information and resources involved in moving a product or service from supplier to customer [3]. The performance of SCM depends on the efficiency of complicated collaboration and integrated management in the whole of supply chain. Information technology, such as database, application software and communication network, has been widely applied in SCM to improve its performance. It has solved the information sharing problem and some ordinary management problems successfully, but there are still a lot of intelligent activities to be improved, such as negotiation, decision and collaboration. Especially in the dynamic business environment, agile capability and flexibility have been the new requirements in today's SCM, and those activities have become the "bottleneck" to affect performance.

This paper explores the application of mobile agent system to deal with the intelligent activities in SCM. It first discusses the development of mobile agent and its application in SCM. With the case of flower trading, we present a mobile agent system for the SCM. It can deal with the negotiation, decision and collaboration intelligently and automatically.

This paper is organized as follows: Section 2 discusses the development of mobile agent and its application in SCM. Section 3 presents a mobile agent system for the case of SCM in flower trading. Section 4 is the conclusion of this paper.

II. DEVELOPMENT OF MOBILE AGENT AND ITS APPLICATION IN SCM

A. Mobile Agent and Its Development

A Mobile Agent (MA) is actually the hybrid of distributed computing technology and agent technology. It is a type of software agents with the features of autonomy, social ability, learning, and the most important feature mobility [4].

A mobile agent consists of three parts: code, state and data. The code is executed when mobile agent migrates to a new platform. The state is the data execution environment of the agent, including the program counter and the execution stack. The data consists of the variables used by the agents, such as knowledge, file identifiers. And there're two primary types of migration of mobile agent: strong migration and weak migration [5]. Strong migration is more complex and is the case where an agent's execution is frozen, migration takes place and then execution is restarted from the very next instruction [5]. Weak migration does not send the agent state, and agent execution restarts from the beginning of the code.

There're already many platforms for the development of mobile agent, such as Aglets, Ajanta, JADE, and Voyager. Due to the features of JADE, we choose it as the mobile agent development platform. JADE is a software platform that provides basic middleware-layer functionalities. It contains two parts: a platform for agent following FIPA standard, and a software package for Java agent development. A significant merit of JADE is that it

This research was supported by National High-tech R & D Program (863 Program) of China (No.2008AA04Z127) and Shanghai Leading Academic Discipline Project (No.B210).
Corresponding author: Weihui Dai.

implements this abstraction over a well-known object-oriented language, Java, providing a simple and friendly API [5]. Figure 1 is the reference architecture of agent platform following the FIPA standard.

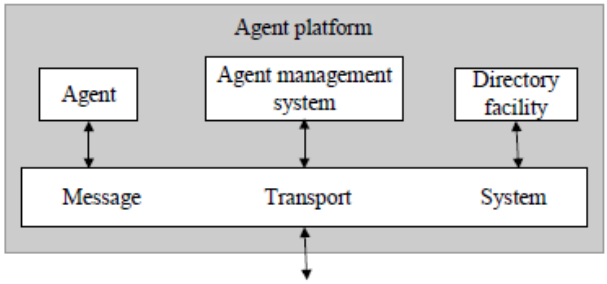


Figure 1. Reference architecture of agent platform following the FIPA standard

JADE platform is composed of agent containers that can be distributed over the network. Agents live in containers which are the Java process that provides the JADE run-time and all the services need for hosting and executing agents. We can consider agents as objects, but agents have some special traits: they are autonomous (i.e. they decide for themselves whether or not to perform an action on request from another agent); they are capable of a flexible behavior; and each agent of a system has its own thread of control [6].

Agent Management System (AMS) and Directory Facilitator (DF) are parts of the platform. There's only one AMS on one platform, which offers the white page service and agent lifecycle service, and maintain the directory of agent identifiers (AID) and the state of agent. DF offers the yellow page service.

B. Application of Mobile Agent in SCM

As mentioned in [8]-[10], there're some advantages when mobile agent is applied in SCM: (1) Mobile agent is the delegate of tasks; (2) Mobile agent can reduce the network load; (3) Mobile agent moves autonomously and asynchronously; (4) Mobile agent facilitates parallel processing; (5) Code shipping rather than data shipping.

However, there're some key problems that have to be considered:

(1) Design pattern of the mobile agent: Because mobile agent itself has size, when it is sent out to fulfill the task, some information is encapsulated. After finishing the task, the result will be added to the mobile agent. Thus the size of mobile agent will increase. Considering the size of mobile agent, the size of information, the number of mobile agents roaming on the net, and the limitation of bandwidth, how to design the mobile agent?

(2) Load balance of nodes: In the supply chain, there's a possibility that a large number of mobile agents roaming on the net, and many agents arrive at the same node at the same time. How does the node agent provide services to the waiting agents efficiently to avoid dead-lock phenomenon happening?

(3) Data synchronization: The market is dynamic and changeful, when mobile agents are moving on the way to the destinations, but the information of the destinations is being changed or has been changed, how do the mobile agents sense the change in time, and how to change the moving route?

(4) Routing planning: When mobile agents are assigned to complex tasks, and at the same time they need to move to lots of destination nodes, how to make a high efficient route to finish the task in short time?

(5) Security problem: How to enforce the security to protect the sensitive information, and how to protect the hosts and agents from attacking?

In [7], there are eight kinds of agent design patterns: Itinerary, Star-Shaped, Branching, Master-Slave, MoProxy, Meeting, Facilitator, and Mutual Itinerary Recording. Here, we choose the branching pattern to be considered in SCM and discuss how to solve the related problems in the following section.

On the Branching pattern, the agent receives a list of agencies to visit and clones itself according to the numbers of agencies in the itinerary. Then, all clones will visit an agency of the received list. Figure 2 shows this pattern.

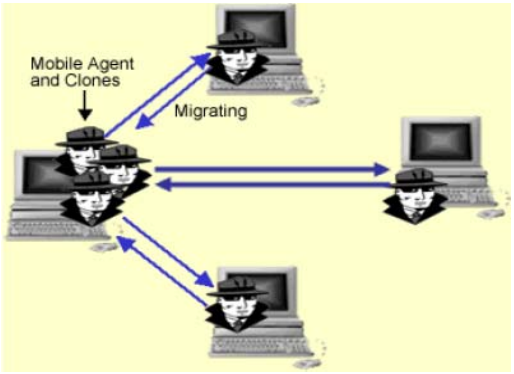


Figure 2. Branching pattern

JADE provides Agent Mobility Service which implements intra-platform mobility. The agent mobility is simply controlled via the method doMove() in the Agent class. Java serialization is used to transmit an agent instance over a network connection by recursively recording the internal member values of the agent object into a byte stream [5].

When mobile agent needs to migrate to another platform, the Inter-Platform mobility Service (IPMS) is provided by JADE. FIPA-ACL messages are used as the transportation medium and these messages are sent between the AMSs of the endpoint platforms [5].

III. MOBILE AGENT SYSTEM FOR SCM IN FLOWER TRADING

A. System Structure

We present a mobile agent system for the case of SCM in flower trading. In the trading, the participants are buyer, seller, supplier and logistics company.

The structure of mobile agent system is designed as Figure 3. It includes four sub-agent systems and an Agency Agent System (AAS). In every sub-agent system, there're three agent modules. For example, modules in the SellerAgent sub-system are: (1) SellerAgent indicates

the special seller, and registers its own information and services on the AMS; (2) Local information serviceAgent is a non-mobile agent module and supplies services to local agents or MAs of other sub-system; (3) Mobile agent module is responsible for managing all the MAs created by local node and maintaining the agents' information table. Every agent module is not a single agent, but a combination of different function agents. Agency agent system works on the web server to manage all the agents' information and all the services.

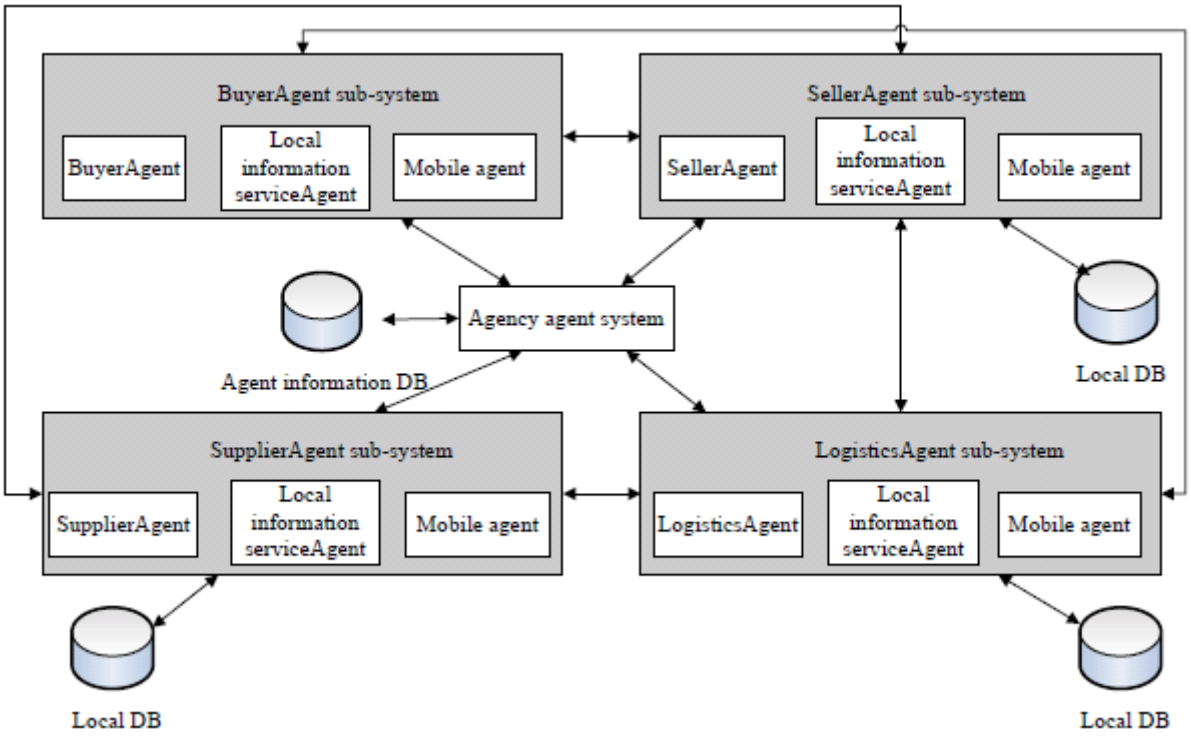


Figure 3. Structure of mobile agent system

During the flower trading processes, agents communicate or even negotiate with others as intelligent entities rather than just sending and receiving information as information reporting tools. So collaboration and interoperation among agents are the essential activities to fulfill the task. Figure 4 is the structure of an agent.

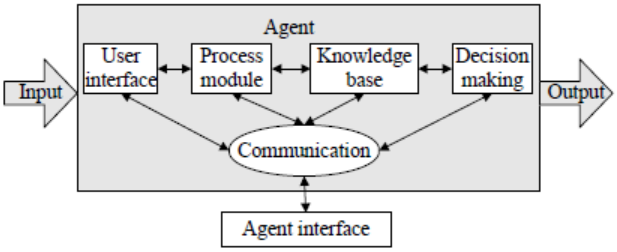


Figure 4. Structure of agent

In this system, buyers can search the flowers which are on market. When the order is confirmed, one or more MAs will be created to carry out the tasks on behalf of buyers. We choose branching as the MAs' design pattern, and the migration type of MAs is weak migration. The basic activities of MA are: create, replicate, migrate, recall, suspend, wake, and destroy. On the Web Server, the agency agent system is configured, which contains the AMS. Every agent must register on the AMS to obtain a unique AID, and they also need to register/deregister their services on the AMS. Therefore, every registered service can be found when there's a service requirement.

B. System Operation

There're some function agents in this system: BuyerAgent, SellerAgent, SupplierAgent, LogisticsAgent, PurchaseAgent, QueryAgent, NegotiationAgent, OrderAgent, ManagerAgent. OrderAgents are mobile agents, which are encapsulated with different information in different process phases and for different participants. The system operation can be described as follows:

(1) All the agents of sub-system register their own information and services on the AAS.

(2) A buyer login the website by the browser, then choose flowers he want, and input some query conditions, such as the time limit of flowers, expected price, amount, delivery date.

(3) According to the order information, BuyerAgent sends the service requirements to the AAS, the agency agent searches whether the required services exist, if there're the right services, it will send the addresses and names of the service providers to the BuyerAgent. After obtaining the addresses of target service providers, it forms a route list for MAs, and MAs migrate to the destinations to query the matched information of flowers.

(4) QueryAgent queries according to the conditions, and then transfers the query result to MAs.

(5) When the buyer and the seller need to negotiate about the price, date and other information, NegotiationAgent will carry out the negotiation according to the negotiation mechanism, and then transfer the result

to MAs.

(6) If there're not enough amount of flowers for the order, SellerAgent will communicate with the SupplierAgent immediately to make supplement.

(7) MA moves back with the negotiation result to BuyerAgent.

(8) The buyer makes a final confirm to choose the right seller.

(9) When the trading is successful, a logistics order is generated, and MA brings this order information to the logistics company. And then the logistics company arranges the delivery.

(10) If the trading fails, both the buyer and the seller will receive the failure inform.

C. Query and Negotiation Mechanism

The flower is a special kind of goods, because flower's lifecycle is short. When considering the query and negotiation activities, agents should obey the following mechanism:

(1) Sellers maintain their own database for all the flowers they sell. Flower's lifecycle is one of the information items (lifecycle = leaving off market time– coming into market time). And the order contains buyer's expected delivery date.

(2) Match according to the date: if the expected delivery date exceeds the flower's leaving off market date, this candidate will be discarded; if the expected delivery date between the flower' s coming into market date and leaving off market date, the relative information will be transferred to MAs.

(3) Match according to the price: if the expected price is lower than all the flowers' price, trading fails. But the flower information of lowest price will be transferred to MAs, in order to give the buyer another chance to make an order. If the expected price is between all the flowers' lowest and highest price, flowers will be sold at the current price.

(4) The flower has discount price, for every seller, the lowest price of the flower is the discount price, and the formula is: lowest price = discount price = original price * (leaving off market time – current time) / (leaving off time– coming into market time).

Because of the special trading scenario, we need to define an ontology in order to keep agents have consistent understanding of the information. The ontology indicates the vocabulary of the symbols used in the content. Both the information sender and the receiver must ascribe the same meaning to these symbols for the communication to be effective. On JADE platform, code of ontology realizing, and code of sending and receiving information, are independent of the content language. In the Flower-trading system, we define a vocabulary named FlowerTradingVocabulary.

With the help of mobile agent system, most of the intelligent activities in SCM can be processed intelligently and automatically.

Ⅳ. CONCLUSION

In dynamic business environment, agile capability and flexibility have been the new requirements in SCM. How to improve the efficiency of intelligent activities has become the "bottleneck" and exploring point in today's SCM. In this paper, we present a mobile agent system to make some of those intelligent activities processed intelligently and automatically. Further researches are expected to improve the agent capability with a perfect knowledge base and enforce its security to protect the sensitive information.

ACKNOLEDGMENT

This research is supported by National High-tech R & D Program (863 Program) of China (No.2008AA04Z127) and Shanghai Leading Academic Discipline Project (No.B210).

REFERENCES

[1] http://en.wikipedia.org/wiki/Supply_chain.

[2] Tarokh, M.J., Bagherzadeh, M., kahani, N., "Supply Chain Coordination Using Role Based Mobile Agent," Service Operations and Logistics, and Informatics, 2006. SOLI '06. IEEE International Conference on 21-23 June 2006, Page(s):322 – 327.

[3] http://en.wikipedia.org/wiki/Supply_chain_management.

[4] http://en.wikipedia.org/wiki/Mobile_agent.

[5] Fabio Bellifemine, Giovanni Caire, Deminic Greenwood, *Developing Multi-Agent Systems with JADE*. John Wiley & Sons Ltd.

[6] Yang Hang, Simon Fong, "Double-agent Architecture for Collaborative Supply Chain Formation," Nov. 2008 Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services.

[7] Zakaria Maamar, Paul Labbé, "Moving vs. Inviting Software Agents: What is the best strategy to adopt?" Communications of the ACM, Volume 46, Issue 7 (July 2003), Pages: 143 – 144.

[8] P.braun, W.Rossak, "Mobile Agents Basic concept, Mobility model and Tracy Toolkit," 2005.

[9] Dongming, Xu, H.Wang, "Multi-agent collaboration for B2B workflow monitoring," Knowledge-Base Systems 15 (2002) 485-491.

[10] Dan Shiao, "Mobile Agent: New Model of Intelligent Distributed Computing," IBM China, October, 2004.

Author Index