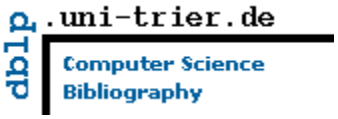
IJCSI

International Journal of Computer Science Issues

Volume 8, Issue 6, No 3, November 2011
ISSN (Online): 1694-0814

www.IJCSI.org

IJCSI proceedings are currently indexed by:



IJCSI Publicity Board 2011

IJCSI Editorial Board 2011

EDITORIAL

In this sixth and last edition of 2011, we bring forward issues from various dynamic computer science fields ranging from system performance, computer vision, artificial intelligence, software engineering, multimedia, pattern recognition, information retrieval, databases, security and networking among others.

Considering the growing interest of academics worldwide to publish in IJCSI, we invite universities and institutions to partner with us to further encourage open-access publications.

As always we thank all our reviewers for providing constructive comments on papers sent to them for review. This helps enormously in improving the quality of papers published in this issue.

Google Scholar reported a large amount of cited papers published in IJCSI. We will continue to encourage the readers, authors and reviewers and the computer science scientific community and interested authors to continue citing papers published by the journal.

It was with pleasure and a sense of satisfaction that we announced in mid March 2011 our 2-year Impact Factor which is evaluated at 0.242. For more information about this please see the FAQ section of the journal.

Apart from availability of the full-texts from the journal website, all published papers are deposited in open-access repositories to make access easier and ensure continuous availability of its proceedings free of charge for all researchers.

We are pleased to present IJCSI Volume 8, Issue 6, No 3, November 2011 (IJCSI Vol. 8, Issue 6, No 3). The acceptance rate for this issue is 32.8%.

IJCSI Editorial Board
November 2011 Issue
ISSN (Online): 1694-0814

www.IJCSI.org

IJCSI Reviewers Committee 2011

• Mr. Markus Schatten, University of Zagreb, Faculty of Organization and Informatics, Croatia
• Mr. Vassilis Papataxiarhis, Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece
• Dr Modestos Stavrakis, University of the Aegean, Greece
• Dr Fadi KHALIL, LAAS -- CNRS Laboratory, France
• Dr Dimitar Trajanov, Faculty of Electrical Engineering and Information technologies, ss. Cyril and Methodius Univesity - Skopje, Macedonia
• Dr Jinping Yuan, College of Information System and Management,National Univ. of Defense Tech., China
• Dr Alexis Lazanas, Ministry of Education, Greece
• Dr Stavroula Mougiakakou, University of Bern, ARTORG Center for Biomedical Engineering Research, Switzerland
• Dr Cyril de Runz, CReSTIC-SIC, IUT de Reims, University of Reims, France
• Mr. Pramodkumar P. Gupta, Dept of Bioinformatics, Dr D Y Patil University, India
• Dr Alireza Fereidunian, School of ECE, University of Tehran, Iran
• Mr. Fred Viezens, Otto-Von-Guericke-University Magdeburg, Germany
• Dr. Richard G. Bush, Lawrence Technological University, United States
• Dr. Ola Osunkoya, Information Security Architect, USA
• Mr. Kotsokostas N.Antonios, TEI Piraeus, Hellas
• Prof Steven Totosy de Zepetnek, U of Halle-Wittenberg & Purdue U & National Sun Yat-sen U, Germany, USA, Taiwan
• Mr. M Arif Siddiqui, Najran University, Saudi Arabia
• Ms. Ilknur Icke, The Graduate Center, City University of New York, USA
• Prof Miroslav Baca, Faculty of Organization and Informatics, University of Zagreb, Croatia
• Dr. Elvia Ruiz Beltrán, Instituto Tecnológico de Aguascalientes, Mexico
• Mr. Moustafa Banbouk, Engineer du Telecom, UAE
• Mr. Kevin P. Monaghan, Wayne State University, Detroit, Michigan, USA
• Ms. Moira Stephens, University of Sydney, Australia
• Ms. Maryam Feily, National Advanced IPv6 Centre of Excellence (NAV6) , Universiti Sains Malaysia (USM), Malaysia
• Dr. Constantine YIALOURIS, Informatics Laboratory Agricultural University of Athens, Greece
• Mrs. Angeles Abella, U. de Montreal, Canada
• Dr. Patrizio Arrigo, CNR ISMAC, italy
• Mr. Anirban Mukhopadhyay, B.P.Poddar Institute of Management & Technology, India
• Mr. Dinesh Kumar, DAV Institute of Engineering & Technology, India
• Mr. Jorge L. Hernandez-Ardieta, INDRA SISTEMAS / University Carlos III of Madrid, Spain
• Mr. AliReza Shahrestani, University of Malaya (UM), National Advanced IPv6 Centre of Excellence (NAv6), Malaysia
• Mr. Blagoj Ristevski, Faculty of Administration and Information Systems Management - Bitola, Republic of Macedonia
• Mr. Mauricio Egidio Cantão, Department of Computer Science / University of São Paulo, Brazil
• Mr. Jules Ruis, Fractal Consultancy, The netherlands
• Mr. Mohammad Iftekhar Husain, University at Buffalo, USA
• Dr. Deepak Laxmi Narasimha, Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia

• Dr. Paola Di Maio, DMEM University of Strathclyde, UK
• Dr. Bhanu Pratap Singh, Institute of Instrumentation Engineering, Kurukshetra University Kurukshetra, India
• Mr. Sana Ullah, Inha University, South Korea
• Mr. Cornelis Pieter Pieters, Condast, The Netherlands
• Dr. Amogh Kavimandan, The MathWorks Inc., USA
• Dr. Zhinan Zhou, Samsung Telecommunications America, USA
• Mr. Alberto de Santos Sierra, Universidad Politécnica de Madrid, Spain
• Dr. Md. Atiqur Rahman Ahad, Department of Applied Physics, Electronics & Communication Engineering (APECE), University of Dhaka, Bangladesh
• Dr. Charalampos Bratsas, Lab of Medical Informatics, Medical Faculty, Aristotle University, Thessaloniki, Greece
• Ms. Alexia Dini Kounoudes, Cyprus University of Technology, Cyprus
• Dr. Jorge A. Ruiz-Vanoye, Universidad Juárez Autónoma de Tabasco, Mexico
• Dr. Alejandro Fuentes Penna, Universidad Popular Autónoma del Estado de Puebla, México
• Dr. Ocotlán Díaz-Parra, Universidad Juárez Autónoma de Tabasco, México
• Mrs. Nantia Iakovidou, Aristotle University of Thessaloniki, Greece
• Mr. Vinay Chopra, DAV Institute of Engineering & Technology, Jalandhar
• Ms. Carmen Lastres, Universidad Politécnica de Madrid - Centre for Smart Environments, Spain
• Dr. Sanja Lazarova-Molnar, United Arab Emirates University, UAE
• Mr. Srikrishna Nudurumati, Imaging & Printing Group R&D Hub, Hewlett-Packard, India
• Dr. Olivier Nocent, CReSTIC/SIC, University of Reims, France
• Mr. Burak Cizmeci, Isik University, Turkey
• Dr. Carlos Jaime Barrios Hernandez, LIG (Laboratory Of Informatics of Grenoble), France
• Mr. Md. Rabiul Islam, Rajshahi university of Engineering & Technology (RUET), Bangladesh
• Dr. LAKHOUA Mohamed Najeh, ISSAT - Laboratory of Analysis and Control of Systems, Tunisia
• Dr. Alessandro Lavacchi, Department of Chemistry - University of Firenze, Italy
• Mr. Mungwe, University of Oldenburg, Germany
• Mr. Somnath Tagore, Dr D Y Patil University, India
• Ms. Xueqin Wang, ATCS, USA
• Dr. Borislav D Dimitrov, Department of General Practice, Royal College of Surgeons in Ireland, Dublin, Ireland
• Dr. Fondjo Fotou Franklin, Langston University, USA
• Dr. Vishal Goyal, Department of Computer Science, Punjabi University, Patiala, India
• Mr. Thomas J. Clancy, ACM, United States
• Dr. Ahmed Nabih Zaki Rashed, Dr. in Electronic Engineering, Faculty of Electronic Engineering, menouf 32951, Electronics and Electrical Communication Engineering Department, Menoufia university, EGYPT, EGYPT
• Dr. Rushed Kanawati, LIPN, France
• Mr. Koteshwar Rao, K G Reddy College Of ENGG.&TECH,CHILKUR, RR DIST.,AP, India
• Mr. M. Nagesh Kumar, Department of Electronics and Communication, J.S.S. research foundation, Mysore University, Mysore-6, India
• Dr. Ibrahim Noha, Grenoble Informatics Laboratory, France
• Mr. Muhammad Yasir Qadri, University of Essex, UK
• Mr. Annadurai .P, KMCPGS, Lawspet, Pondicherry, India, (Aff. Pondicherry Univeristy, India
• Mr. E Munivel , CEDTI (Govt. of India), India
• Dr. Chitra Ganesh Desai, University of Pune, India
• Mr. Syed, Analytical Services & Materials, Inc., USA

• Mrs. Payal N. Raj, Veer South Gujarat University, India
• Mrs. Priti Maheshwary, Maulana Azad National Institute of Technology, Bhopal, India
• Mr. Mahesh Goyani, S.P. University, India, India
• Mr. Vinay Verma, Defence Avionics Research Establishment, DRDO, India
• Dr. George A. Papakostas, Democritus University of Thrace, Greece
• Mr. Abhijit Sanjiv Kulkarni, DARE, DRDO, India
• Mr. Kavi Kumar Khedo, University of Mauritius, Mauritius
• Dr. B. Sivaselvan, Indian Institute of Information Technology, Design & Manufacturing, Kancheepuram, IIT Madras Campus, India
• Dr. Partha Pratim Bhattacharya, Greater Kolkata College of Engineering and Management, West Bengal University of Technology, India
• Mr. Manish Maheshwari, Makhanlal C University of Journalism & Communication, India
• Dr. Siddhartha Kumar Khaitan, Iowa State University, USA
• Dr. Mandhapati Raju, General Motors Inc, USA
• Dr. M.Iqbal Saripan, Universiti Putra Malaysia, Malaysia
• Mr. Ahmad Shukri Mohd Noor, University Malaysia Terengganu, Malaysia
• Mr. Selvakuberan K, TATA Consultancy Services, India
• Dr. Smita Rajpal, Institute of Technology and Management, Gurgaon, India
• Mr. Rakesh Kachroo, Tata Consultancy Services, India
• Mr. Raman Kumar, National Institute of Technology, Jalandhar, Punjab., India
• Mr. Nitesh Sureja, S.P.University, India
• Dr. M. Emre Celebi, Louisiana State University, Shreveport, USA
• Dr. Aung Kyaw Oo, Defence Services Academy, Myanmar
• Mr. Sanjay P. Patel, Sankalchand Patel College of Engineering, Visnagar, Gujarat, India
• Dr. Pascal Fallavollita, Queens University, Canada
• Mr. Jitendra Agrawal, Rajiv Gandhi Technological University, Bhopal, MP, India
• Mr. Ismael Rafael Ponce Medellín, Cenidet (Centro Nacional de Investigación y Desarrollo Tecnológico), Mexico
• Mr. Supheakmungkol SARIN, Waseda University, Japan
• Mr. Shoukat Ullah, Govt. Post Graduate College Bannu, Pakistan
• Dr. Vivian Augustine, Telecom Zimbabwe, Zimbabwe
• Mrs. Mutalli Vatila, Offshore Business Philipines, Philipines
• Mr. Pankaj Kumar, SAMA, India
• Dr. Himanshu Aggarwal, Punjabi University,Patiala, India
• Dr. Vauvert Guillaume, Europages, France
• Prof Yee Ming Chen, Department of Industrial Engineering and Management, Yuan Ze University, Taiwan
• Dr. Constantino Malagón, Nebrija University, Spain
• Prof Kanwalvir Singh Dhindsa, B.B.S.B.Engg.College, Fatehgarh Sahib (Punjab), India
• Mr. Angkoon Phinyomark, Prince of Singkla University, Thailand
• Ms. Nital H. Mistry, Veer Narmad South Gujarat University, Surat, India
• Dr. M.R.Sumalatha, Anna University, India
• Mr. Somesh Kumar Dewangan, Disha Institute of Management and Technology, India
• Mr. Raman Maini, Punjabi University, Patiala(Punjab)-147002, India
• Dr. Abdelkader Outtagarts, Alcatel-Lucent Bell-Labs, France
• Prof Dr. Abdul Wahid, AKG Engg. College, Ghaziabad, India
• Mr. Prabu Mohandas, Anna University/Adhiyamaan College of Engineering, india
• Dr. Manish Kumar Jindal, Panjab University Regional Centre, Muktsar, India

• Mr. Neetesh Gupta, Technocrats Inst. of Technology, Bhopal, India
• Ms. A. Lavanya, Manipal University, Karnataka, India
• Ms. D. Pravallika, Manipal University, Karnataka, India
• Prof. Vuda Sreenivasarao, St. Mary's college of Engg & Tech, India
• Prof. Ashutosh Kumar Dubey, Assistant Professor, India
• Mr. Ranjit Singh, Apeejay Institute of Management, Jalandhar, India
• Mr. Prasad S.Halgaonkar, MIT, Pune University, India
• Mr. Anand Sharma, MITS, Lakshmangarh, Sikar (Rajasthan), India
• Mr. Amit Kumar, Jaypee University of Engineering and Technology, India
• Prof. Vasavi Bande, Computer Science and Engneering, Hyderabad Institute of Technology and Management, India
• Dr. Jagdish Lal Raheja, Central Electronics Engineering Research Institute, India
• Mr G. Appasami, Dept. of CSE, Dr. Pauls Engineering College, Anna University - Chennai, India
• Mr Vimal Mishra, U.P. Technical Education, Allahabad, India
• Dr. Arti Arya, PES School of Engineering, Bangalore (under VTU, Belgaum, Karnataka), India
• Mr. Pawan Jindal, J.U.E.T. Guna, M.P., India
• Prof. Santhosh.P.Mathew, Saintgits College of Engineering, Kottayam, India
• Dr. P. K. Suri, Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, India
• Dr. Syed Akhter Hossain, Daffodil International University, Bangladesh
• Mr. Nasim Qaisar, Federal Urdu Univetrsity of Arts , Science and Technology, Pakistan
• Mr. Mohit Jain, Maharaja Surajmal Institute of Technology (Affiliated to Guru Gobind Singh Indraprastha University, New Delhi), India
• Dr. Shaveta Rani, GZS College of Engineering & Technology, India
• Dr. Paramjeet Singh, GZS College of Engineering & Technology, India
• Prof. T Venkat Narayana Rao, Department of CSE, Hyderabad Institute of Technology and Management , India
• Mr. Vikas Gupta, CDLM Government Engineering College, Panniwala Mota, India
• Dr Juan José Martínez Castillo, University of Yacambu, Venezuela
• Mr Kunwar S. Vaisla, Department of Computer Science & Engineering, BCT Kumaon Engineering College, India
• Prof. Manpreet Singh, M. M. Engg. College, M. M. University, Haryana, India
• Mr. Syed Imran, University College Cork, Ireland
• Dr. Namfon Assawamekin, University of the Thai Chamber of Commerce, Thailand
• Dr. Shahaboddin Shamshirband, Islamic Azad University, Iran
• Dr. Mohamed Ali Mahjoub, University of Monastir, Tunisia
• Mr. Adis Medic, Infosys ltd, Bosnia and Herzegovina
• Mr Swarup Roy, Department of Information Technology, North Eastern Hill University, Umshing, Shillong 793022, Meghalaya, India
• Mr. Suresh Kallam, East China University of Technology, Nanchang, China
• Dr. Mohammed Ali Hussain, Sai Madhavi Institute of Science & Technology, Rajahmundry, India
• Mr. Vikas Gupta, Adesh Instutute of Engineering & Technology, India
• Dr. Anuraag Awasthi, JV Womens University, Jaipur, India
• Dr. Mathura Prasad Thapliyal, Department of Computer Science, HNB Garhwal University (Centr al University), Srinagar (Garhwal), India
• Mr. Md. Rajibul Islam, Ibnu Sina Institute, University Technology Malaysia, Malaysia
• Mr. Adnan Qureshi, University of Jinan, Shandong, P.R.China, P.R.China
• Dr. Jatinderkumar R. Saini, Narmada College of Computer Application, India

TABLE OF CONTENTS

A Tunable Checkpointing Algorithm for the Distributed Mobile Environment

Sungchae Lim

Dept. of Computer Science, Dongduk Women's University
Seoul, 136-714, South Korea

Abstract

The aim of a distributed checkpointing algorithm is to efficiently restore the execution state of distributed applications in face of hardware or software failures. Originally, such algorithms were devised for fixed networking systems, of which computing components communicate with each other via wired networks. Therefore, those algorithms usually suffer from heavy networking costs coming from frequent data transits over wireless networks, if they are used in the wireless computing environment. In this paper, to reduce usage of wireless communications, our checkpointing algorithm allows the distributed mobile application to tune the level of its checkpointing strictness. The strictness is defined by the maximum rollback distance (MRD) that says how many recent local checkpoints can be rolled back in the worst case. Since our algorithm have more flexibility in checkpointing schedule due to the use of MRD, it is possible to reduce the number of enforced local checkpointing. In particular, the amount of data transited on wirelesses networks becomes much smaller than in earlier methods; thus, our algorithm can provide less communication cost and shortened blocking time.

Keywords: *Mobile networks, distributed application, rollback, recovery, distributed checkpointing.*

1. Introduction

During the past decades, there have been dramatic advances in mobile networks and mobile devices. In particular, the fast spreading usage of smart phones is likely to yield demands for sophisticated distributed applications across multiple mobile devices [1, 2]. During the run-time of such a distributed application, its cooperating application processes (APs) work in parallel and data are usually transited between APs to share application contexts. In this situation, failure on a single AP or hardware device could cause a serious problem in the whole distributed application and thus it may roll back the application's processing state to the initial one in the worst case. To prevent a whole cancelation of the processing result, checkpoint records are created to log intermediate execution results. The recovery procedure after abrupt failure builds a consistent state of an application from the checkpoint data, and resumes the interrupted application from that state. This can reduce undesirable loss of application process

To make the distributed application robust and recoverable against failure, many works are done for the computing environments where the distributed application seems to be executed in the wired fixed networks [4, 6, 7, 10, 11, 13]. When checkpointing algorithms of those earlier works are applied to distributed applications running on wireless networked, they suffer from high cost for sending checkpoint data via warless connections. Since data transit over wireless networks is more costly and unstable, compared with that over wired networks, many researches focus on reduction of wireless data transit in the case of the checkpointing scheme for wireless computing environment [5, 8, 9,10, 12].

In the paper, we also propose a distributed checkpointing scheme suitable for distributed applications running on the mobile computing environment. We here introduce two key ideas of the maximum rollback distance (MRD) and the logging agent running on the MSS (Mobile Support Station). The logging agent is a software agent running on MSS, which is responsible for making local checkpoints at the request of its associated AP's requests and maintaining at least one consistent global checkpoint. To save the cost for maintaining such a global checkpoint, the agent can do some logging activities without any requests from associated AP. For this, the logging agent securitizes messages arriving on its MSS and communicates with other logging agents for synchronization of checkpointing.

If the rule of checkpointing synchronization is too strict, enforced local checkpoints are frequently created. Sine enforced local checkpoint request costly data transit in wireless network lines; it is needed to make the synchronization rule more flexible. In this notion, we introduce the MDR for each distributed application. On the other hand, the MDR is a run-time parameter for a distributed application, saying how many local checkpoint of a given AP can be rolled back in the worst case. With a MDR properly set to a value, a significant flexibility is available at the time local checkpoints are synchronized in order to create a new global checkpoint. Due to the tunable level of checkpointing synchronization using MDR, the logging agents participating in a distributed application can reduce the number of enforced local checkpointing and costly message transit over wireless networks.

The rest of this paper is organized as follows. In Section 2, we describe some backgrounds regarding the meaning of global consistency of distributed checkpoints, the assumed mobile network, and the previous works. Then, we propose a new efficient distributed checkpointing scheme in Section 3, and discuss the performance characteristics of our scheme in Section 4. Lastly, we conclude this paper in Section 5.

2. Backgrounds

2.1 Global Consistent State

The GCS (Global Consistent State) of distributed applications was formally defined by Lamport [17]. According to that definition, processing of a distributed application can be modeled by three types of events such as the message sending event, the message receiving event, and the computation event. Each AP participating in a distributed event can do the computation event to proceed with its processing state and communication with other participant APs through message sending/receiving events.

In this event model, a set of events meeting the GCS can be captured using the relation "happen-before" drawn on events. In [17], the "happen-before" relation (HBR) is as follows.

[Definition of HBR] If it is the case that *e1* "*happen-before*" *e*2, then either of the following conditions should be true.

i) Both *e1* and *e2* occur in the same AP and *e1* precedes *e2* in timing sequence.

ii) There are a message m and two APs of *p1* and *p2* such that *p1* sends (event *e1*) message *m* to *p2* and *p2* receives (event *e2*) it.

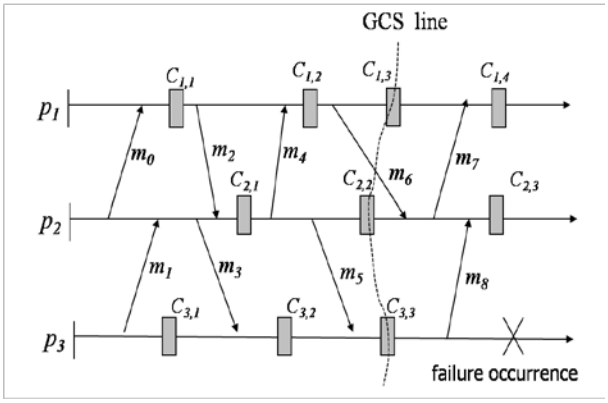


Fig. 1. An example of GCS: a failure arises at *p3*.

Owing to the transitive property of the HBRs, we can give a partial order to the events of a distribute application, even though there is no common clock shared by its participant AP's.

The GCS of a distributed application is defined based on the HBR above. Let us take a snapshot of execution state of a distributed application at a particular time, and let *S* be that snapshot, which is a set of the events having arisen in the application. Let *G* be a subset of *S*. In this case, *G* is said to be in a GCS if the following condition is satisfied; for every event *e'* in *G*, if there is *e* in *S* such that *e* "happen-before" *e'*, then *e* should be also an event in *G*. In other words, for every event of *G*, its causal events should be found in *G*. Since all the causal events are contained in G, it may be possible to obtain the same execution results of G, if we redo the events of G from it begging time. Based on this idea, we can recover any intermediate execution state of any failed application if its any GCS execution snapshot is available.

To have execution snapshots, checkpointing schemes are commonly used for saving local execution state of individual AP's. Fig.1 shows an example where distributed checkpoint is performed by three AP's, *p1*, *p2*, and *p3*. In the figure, the blacked rectangle of $C_{i,k}$ represents the *k*-th local checkpoint made by AP p_i. The local checkpoint of $C_{i,k}$ is made to save the computational computation state of p_i and the message sent to other AP's after the creation time of $C_{i,k-1}$.

Suppose that an application failure arise at *p3*. as in Fig. 1. At this moment, the set of local checkpoints preserving the GCS are that inside the GCS line of the figure. That is, the latest CGS state is composed of $C_{1,3}$, $C_{2,2}$, and $C_{3,3}$. As the message sending event of m8 is not saved in any local checkpoint, its message receiving event cannot be include a GCS. Therefore, $C_{2,2}$ is rolled back, and $C_{1,3}$ is also rolled back because $C_{1,3}$ contains the message-receiving event of m_7 saved in $C_{2,2}$. As a result, the local

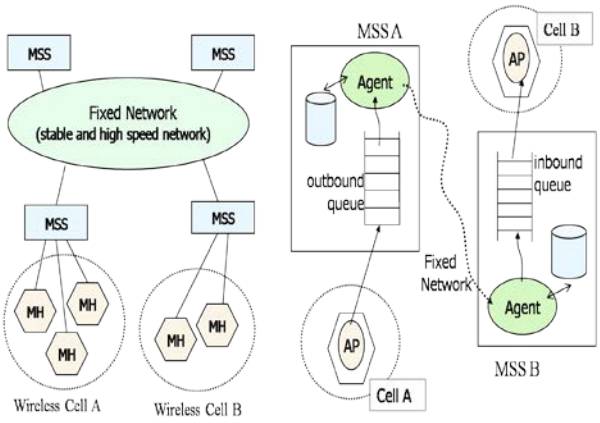
checkpoints on the GCS line will be used to recover the failed application.

In the example of Fig. 1, the latest local checkpoints of a GCS are the same as $C_{1,3}$, $C_{2,2}$, and $C_{3,3}$ is used as a latest consistent global checkpoint. This choice of such a consistent global checkpoint is done by a recover algorithm initiated in the presence of failure. During the recovery phase, such a latest consistent global checkpoint is found and then the disrupted distributed application is restarted. In the case of Fig. 2, the associated AP's will restore their computational state using the data saved in the local checkpoints of $C_{1,3}$, $C_{2,2}$, and $C_{3,3}$, respectively, and *p1* will send the lost message m_6 to *p2* again. Since checkpoint records bookkeep the serial numbers of messages transferred among AP's, this message resending is possible.

2.2 Assumed Mobile Network

In general, the mobile network is comprised of mobile hosts (MH's), mobile support stations (MSS's), and the fixed networks interconnecting the MSSs [1, 2, 8]. The network architecture is shown in Fig. 2(a), where there are two wireless cells and MH's can make wireless network connections within its wireless cell. Since the MH can hop among wireless cells, the MSS's have to update the list of MH's under control for seamless hand-offs. Each AP can be identified by unique process id within its hosting MH. Of cause, the MH is also uniquely identified in the global network environment.

Consider a situation where a MH x in MSS A sends a network message m to an MH y in MSS B. The message m from x is queued into an outbound queue of MSS A and then it is delivered to the counterpart MSS B via the fixed network. Consecutively, message m is entered into the inbound queue of MSS B for the delivery towards y.
On the top of the traditional architecture of Fig. 2(a), we assume that an agent program executes on each MSS for doing checkpoint-related activities. That is, it is assumed that the agent program makes accesses the two message queues of the MSS, in which outbound or inbound network messages are temporarily stored waiting for their delivery to target AP. Fig. 2(b) depicts the assumed architecture with logging agents. In the example of Fig. 2(b), an MH in cell A sends a network message m to other MH staying in cell B. In this case, the logging agent in MSS A dequeues message m and then appends some checkpoint-related data to m before it sends m to the logging agent of MSS B. Correspondingly, the logging agent of MSS B deletes the



(a) Traditional network architecture. (b) Assumed network architecture.

Fig. 2. Architecture of the assumed mobile network.

appended data from m before it inputs m into the inbound queue towards the destination MH. During this message transit time, the logging agent's can make checkpoint records in the disk storage installed in the MSS's. Using the logging agent, we can reduce the checkpoint cost and improve the flexibility of the consistent global checkpoints. The more details about the logging agent are described in Section 3.

2.3 Earlier Works

Checkpointing schemes for distributed applications can be roughly categorized into the synchronized schemes and the asynchronized schemes. In the synchronized schemes, when a AP requests a checkpoint, actions for making a consistent global checkpoint are performed such that the newly created global checkpoint includes the current execution state of the checkpoint-requesting AP. From this, the checkpoint-requesting AP can make its crucial results of execution robust to any failure. For such checkpointing, the checkpoint requester AP is blocked until all the causal events of the checkpoint-requested events are saved in the local checkpoint records of the participant AP's. Owing to such creation of a global checkpoint, most of execution results can be restored in the present of failure.

However, because creation of a global checkpoint needs a number of message deliveries and requires some enforced checkpointing of other participant AP's, this scheme suffer from a high network cost and long delay time for checkpointing. Especially, such shortcomings become more serious in the case where the AP's are ruing in mobile network environment [1, 2, 8, 12].

Meanwhile, the asynchronized scheme does not enforce the creation of a consistent global checkpoint at every checkpoint request time. Instead, during the recovery phase a latest consistent global checkpoint is found from the casual dependency of APs' events saved in disturbed local checkpoints. By examining the casual dependency among the previous local checkpoints, the recovery algorithm picks a most recent consistent global checkpoint for recovery. Since each AP can make its local checkpoint in asynchronized manner, this scheme is apt to have a problem of many cascaded rollbacks of local checkpoints, so-called domino effect [1, 8, 17]. In a worst case, the whole execution results of any distributed application can be cancelled because of the domino-effect. In addition, the asynchronized schemes have more restarting overheads, compared to the synchronized schemes. This is because the asynchronized scheme has to collect the whole information from scatted local checkpoints in order to find a consistent global checkpoint. From these reasons, the synchronized scheme is preferred in earlier time.

However, when it comes to the mobile network environment, the synchronized scheme is more feasible because of less network connectivity and more consideration of instability of mobile devices of that application environment. Among the asynchronized schemes, in particular, the message-induced checkpointing scheme [8, 9] is regarded to be a good alternative solution in the mobile network environment. This is because the message-induced scheme can eliminate the possibility of domino effect sin a very simple manner. By forcing AP's to make local checkpoints depending on the message-receiving events, this scheme can set some boundaries on rollbacked local checkpoints

3. Proposed Method

3.1 Motivations

Although the message-induced scheme is useful to avoid the domino effect, it has a severe problem in that the time of checkpoint creations is determined without active involvement of participant AP's. This problem can be easily understood by viewing the used mechanism of the message-induced scheme. The scheme uses two different execution states of the AP, that is, SEND and RECEIVE states. The SEND and RECIVED states are set while the message-sending and message-receiving events are successively arising, respectively. A new checkpoint is compulsorily created at the time when a network message arrives at a particular AP with the execution state of SEND. Since the time of checkpoint creations in an AP is inactively determined depending on the arrivals of message-receiving events, the created checkpoints would not reflect application's semantics. Otherwise, if we want application's semantics-aware checkpointing, that checkpointing time of checkpoint can be chosen by considering the critical points of processed application. In other words, checkpoints have to be made when some critical executions or expensive processes are done. Such semantics-aware checkpointing is possible only when the AP can actively request checkpointing. Note that the user can also issue checkpoint requests via its AP.

The lack of semantics-awareness of the message-induced scheme may have poor performance. When this scheme is used for distributed mobile applications requiring a lot of message transit, a large number of non-meaningful checkpoints can be made. This can result in frequent checkpointing and creations of obsolete checkpoints. In addition, the message-induced scheme has no mechanism to actively create consistent global checkpoints. Therefore, in the case that an AP wants to make its critical execution results persistent, there is no way for that. Whether or not the execution results are saved into a consistent global checkpoint relies on the existence of an appropriate pair of message-receiving and message-sending events.

To solve such problems of lack of semantics-awareness in checkpointing time and its defective mechanism for global checkpointing, we propose a new checkpoint scheme based on a combination of the logging agent and the R-distance

3.1 Data Format

First, we describe the data format of the network message, which is represented by **M** below. In the followings, the message sender AP is denoted by P_i and the total number of AP's joining the distributed application by N, respectively. The fields of **M** are seven in all. Among them, the last three fields are not used by the AP. These fields exist for containing control data of checkpointing purpose and they are visible only to the logging agent. Meanwhile, the first four are for containing application data needed by the AP.

- **M**.*type*: Message type
- **M**.*sender*: *Id* of P_i
- **M**.*recever*: *Id* of the counterpart AP
- **M**.*data*: Application data sent to the counterpart AP
- **M**.*ap*[*1,2,…,N*]: Ids of AP's joining this application.
- **M**.*serial*[*1,2,…,N*]: Serial numbers of the local checkpoints already made by **M**.*ap*[*1, 2, …N*].
- **M**.*dep_vec*[*1,2,…,N*]: Checkpoint dependency vector

To make the checkpoint-related fields invisible to the entire participant AP's, the logging agent appends these fields at the tail of any message *m* received from its AP before it sends message *m* to other logging agent. Then, the receiver logging agent will delete these fields from *m* before *m* is sent to the destination AP. The dependency vector used for finding a consistent global checkpoint is the same as that proposed in the earlier literatures [7, 8, 9, 11]. For space limitation, the details of use of the dependency vector are referred to the literatures.

In turn, we describe the data format of the checkpoint record managed by the logging agent. The fields below are ones existing in the checkpoint record. When an AP initiates or joins a distributed application, a logging agent of the AP creates a new checkpoint record for saving transferred network messages and checkpoint-related data until the next checkpointing time. At the creation time, the checkpoint record is manipulated in an area of main memory, and then it is written into a stable storage at the next checkpointing time. In the followings, the owner AP of the checkpoint record is represented by P_i, and the total number of AP's joining the distributed application by *N*.

- **REC**.*id*: Id of P_i.
- **REC**.*serial*: Current checkpoint serial number
- **REC**.*r_distance*: R-distance of this application.
- **REC**.*ap*[*1,..,N*]: Ids of participant AP's
- **REC**.*serial*[*1,..,N*]: Serial numbers of the local checkpoints already made by **REC**.*ap*[*1, 2, …N*].
- **REC**.*dep_vec*[*1,..,N*]: checkpoint dependency vector
- **REC**.*message*[]: Messages sent by P_i after the last checkpointing time
- **REC**.*prev_rec*: Disk address to the previous checkpoint record

Fig. 3. logging agent algorithm for handling a message-receiving event.

As known from the above, at the first four fields the checkpoint record saves the *id* of the owner AP, the serial number of the current checkpoint record, the given R-distance, and *id*s of the participant AP's. And, the next two fields are used for bookkeeping the information about created local checkpoint serials and dependency vector.

To log all the network messages sent by P_i until the next checkpointing time, we use the field of *message[]*. Since all the network messages sent to other AP's are logged in that field, messages resending can be done during the recovery phase. Since more than one checkpoint record are created o be created for the same application while the application executes, they are chained for the fast access during the recovery time. The last field is used for that purpose, that is, it saves the disk address to the very previous checkpoint record stored in the disk.

3.2 Tunable Checkpointing

The semantics-aware checkpointing requires that local checkpoints be made according to the determination to importance of the current execution state. Here, the expensive processing is some actions whose loss causes many additional network communications or computational overheads. Such expensive processing is according to application semantics, and thus only the involved AP is responsible for its determination. For that reason, the capability of creating a global checkpoint by the AP is needed, as supported in the synchronized checkpoint scheme. However, such capability inevitably results in a high network cost and long blocking-time when the protocol of the previous synchronized checkpointing schemes is applied to the mobile computing environment.

USED DATA: *R* /* current checkpoint record of P_i */

When a message *m* arrives at C_i from P_i

1. **begin**
2. **if** (*m* is for requesting a checkpoint creation) **then**
3. Save the content of *R* into disk space to make a local checkpoint with the serial number of *R*.serial.
4. *R'* ← GetGCSRec(*R*). /* get a latest checkpoint record of a GCS */
5. **if** (*R'*= **nil**)
6. Call the routine *GreateGCS*(*R*).
7. **endif**
8. Create a new checkpoint record with the serial number of *R*.serial + 1.
9. Send a messages notifying the creation of a new local checkpoint of P_i.
10. Send a response message of checkpointing to P_i .
11. **else** /* *m* contains application data sent to other AP */
12. Append checkpoint-related fields to *m* and send it to the counterpart logging agent.
13. **endif**
14. **end.**

When a message *m* arrives at C_i from other logging agent

15. **begin**
16. **if** (*m* is an application data message toward P_i)
17. Call the routine *UpdateChptRec*(*m* , *R*).
18. Remove some checkpoint-related fields form *m* and sent it to P_i .
19. **else if** (*m* is for notifying creation of a new remote checkpoint)
20. *UpdateChptRec*(*m*, *R*).
21. **else** /* *m* is for requesting P_i 's checkpointing */
22. Make a message for requesting an enforced checkpointing and set it to P_i .
23. **endif**
24. **end**.

Fig. 3. logging agent algorithm for handling a message-receiving event.

To solve such a problem, we introduce the notion of the recovery distance (R-distance) for the distributed application. The R-distance indicates the worst-case number of rollbacked checkpoints in the presence of failure. If its value is *d*, then the latest *d* -1 local checkpoints can be rolled back at the worst-case. For example, if its value is equal to three, then the latest two checkpoints can be rolled back with respect to each participant AP. In the same way, if its value is one, our scheme will work identically with an earlier synchronized checkpoint scheme, where every checkpoint request results in creation of a new consistent global checkpoint. If the current application is in a very mission critical state, then the application initiator AP can set the R-distance to a small one. Additionally, if an AP really wants to make a global checkpoint for a distributed application with R-distance *d*, it can do that by issuing *d* local checkpoint requests successively.

The R-distance value is determined and assigned to every distributed application at its beginning point, and the global checkpoints are made in flexible manner, while preserving the given R-distance. In our scheme, the enforced global checkpoint is issued by only the logging agent and the necessity of such enforced checkpointing is also decided by the logging agent. The AP just issues a request for creating its local checkpoint by reflecting application semantics.

The algorithm of Fig. 3 shows the way a logging agent works at the time when a network message *m* arrives at the logging agent. In the algorithm, the logging agent receiving message *m* is denoted by C_i, and the AP checkpointed by C_i is denoted by P_i. The message *m* can be one from P_i or any other logging agent. Since all the messages sent to an AP are relayed by logging agent's, every message from AP's other than P_i arrives at C_i via the logging agent's.

The steps of lines 1–14, are executed if C_i receives a message from P_i. In this case, C_i first checks if message *m* is for requesting a creation of P_i's local checkpoint. If that is true, the steps of lines 3-9 are performed to make a new local checkpoint, preserving the R-distance of the distributed application. For this, C_i calls the routine *GetGCSRec*() to get the lasts checkpoint record of a GCS. Then, the record's serial number is compared with the that of the newly created local checkpoint. If preservation of the R-distance constraint is not possible, then the routine *CreateGCS*() is executed to make a new consistent global checkpoint as in line 6. Otherwise, if the R-distance is preserved, then C_i just saves the current checkpoint record and send a response message back to P_i for notifying successful checkpointing. On the other hand, if *m* is a pure application data message, then the message is delivered to the counterpart logging agent managing the message receiver AP. At that time, some fields used for checkpointing are appended to the original *m*.

First, we describe the data format of the network message, which is represented by **M** below. In the followings, the message sender AP is denoted by P_i and the total number of AP's joining the distributed application by *N*, respectively. The fields of **M** are seven in all. Among them, the last three fields are not used by the AP. These fields exist for containing control data of checkpointing purpose and they are visible only to the logging agent. Meanwhile, the first four are for containing application data needed by the AP.

The rest steps in lines 15-24 of Fig. 3 are ones to be performed when C_i receives *m* from other logging agent, say C_j. If *m* is for sending application data to P_i, then it is sent to P_i, after some piggybacking fields are deleted from *m*. Of course, to save the checkpoint-related data the routine *UpdateChptRec*() is called in line 17. If message *m* is not for sending pure application data, it is either for notifying a new checkpoint creation in the side of C_j or for forcing P_i to create a new local checkpoint. In the former case, C_i just updates the current checkpoint record for reflecting the advance of the remote checkpoint serial number and other changes of the distributed application. Since M has no application data in itself, further message delivery is not needed. In the latter case, a new message forcing P_i to make its local checkpoint is sent to P_i as in line 22. In the response of that message, P_i will send a message for checkpoint creations, and then line 3 is executed later.

The main advantages of our checkpointing scheme in Fig. 3 are two-fold. First, based on the concept of R-distance, the average cost for creating a consistent global checkpoint can be reduced, because creation of the global checkpoint can be delayed within the R-distance. Additionally, if no global checkpoint is found within the R-distance, then our scheme estimate the costs of global checkpoints within R-distance in routine *CreateGCS*(). Those features differentiate our checkpointing protocol from others used for global checkpointing in the earlier schemes, which only have to create a global checkpoint containing the latest local checkpoint without any consideration of its creation cost. With our flexibility and cost estimation in global checkpoint time, we can reduce the average cost for making a global checkpoint.

Second, the use of the logging agent can reduce the amount of checkpoint-related data transferred between MSS's and MH's. Since those additional data for checkpoint is always needed for tracking application's execution state,

```
Algorithm: Routine GetGCSRec(R)
INPUT:  Pi's checkpoint record R
OUTPUT: checkpoint record having a GCS

1.  A[1] ← R.
2.  for i =2 to R.r_distance
3.    A[i] ← ReadChptRec(A[i-1].previous). /* reading of the
      previous checkpoint records */
4.  endfor
5.  p_num ← number of AP's saved in R.ap[].
6.  for i = 1 to R.r_distance
7.     cgs_exist ← yes.
8.     for j = 1 to p_num
9.        if ( A[i].chpt_dep_vec[j] > R.serial[j] )
10.          cgs_exist = no.
11.    endfor
12.    if( cgs_exist = yes ) return A[j].
13. endfor
14. return nil.  /* no GCS checkpoint record */
```

Fig. 4. Algorithm for routine *GetGCSRec*().

overheads for sending them are unavoidable. If the communication overheads become too large on wireless networks, that could be a severe bottleneck in the execution of the distributed applications. Against such a problem, we adopt the logging agent so that most of the additional checkpoint-related data are visible only in the network messages transferred within logging agent's wired fixed networks. Since the communication cost in the fixed networks of logging agent's is very lower than in wireless communication, we can reduce the overall network cost due to the use of logging agents.

3.2 Detailed Algorithms

In Fig. 3, we outlined the proposed algorithm of the logging agent. In that algorithm, we omitted details of the routines called by the logging agent in Fig. 3. Here, we present the detailed algorithms of the routines. Besides those routines of Fig. 3, other routines used by the AP are also needed for our checkpointing scheme. For instance, we need the routines for message sending/receiving, requesting a local checkpoint, and processing a checkpoint enforcement message. As the algorithms for those AP routines are not distinctive from ones previously proposed in [8, 9] and they can be conjectured from the algorithm of the logging agent, we do not present them in this paper.

Fig. 4 depicts the algorithm of routine *GetGCSRec*() used to get a latest checkpoint record of P_i being in a GCS. Here, P_i is the AP whose checkpoint record is *R* of this routine. In lines 1-4, the routine reads the previous checkpoint records into the memory areas of *A*[2], .., *A*[*R*.r_distance] and the current checkpoint record into *A*[1], respectively. For this, the backward pointers chaining the disk-resident checkpoint records are used for fast accesses. Using the dependency vector and the serial local checkpoint numbers saved in *A*[1, …, *R*.r_distance], this routine finds a latest GCS.

The algorithm of routine *GetGCSRec*() is based on the concept of the local checkpoints dependency among different AP's. This is represented by the dependency vector saved in the checkpoint record field of *dep_vec[]* of Fig. 3. The use of dependency vector is common in the global checkpoint schemes [5, 6, 9, 10, 11]. The proof on the usefulness of the dependency vector is also referred to these researches.

We also use the dependency vector for deterring a collection of local checkpoints with a GCS, that is, a consistent global checkpoint. This routine compares the dependency vector saved in R the current serial number of the latest checkpoint records of other AP's. The latest checkpoint serials are found in the data structure of *serial[]* in *R*. With the comparison, this routine can finds the latest local checkpoint whose dependency are checked that is not dependent on the events that have not been check-pointed by counterpart AP's.

In lines 6-13, the logging agent decides whether or not the current creation of a local checkpoint supports the R-distance preservation. If its preservation is not possible due to the current checkpoint request, the routine *CreateGCS*() is called as in line 6 of Fig. 3.

Fig. 5 depicts the algorithm of routine *CreateGCS*(). This routine also reads the previous checkpoint records into *A*[] for fast manipulation. Then, in lines x-x the routine computes the global checkpoint costs with respect to the local checkpoints represented by *A*[]. Here, the costs are assessed by the number of remote local checkpoints to be created for yielding a GCS, plus the distance of the chosen local checkpoint from the current checkpoint time. That is performed in lines 8-15 of Fig. 5. Based on the estimation, the forced GCS line is determined by favorably choosing a local checkpoint with the smallest costs as in line 16.

To make the chosen local checkpoint, denote by *A*[*s*] in the algorithm of Fig. 5, be a global consistent one, messages for requesting enforcement of checkpointing in other AP's are sent to the involved logging agent's. Then, the logging agent's will enforce its AP to create the local checkpoint. When all the response messages are gathered, this routine returns. ,

The routine *UpdateChptRec*() of Fig. 6 is for modifying the checkpoint record in accordance with message arrivals. In line 2, the routine check if m is an outbound message

Algorithm: Routine *CreateGCS(R)*
INPUT: *Pi's* checkpoint record *R*

1. *A*[1] ← *R*.
2. **for** *i* =2 **to** *R*.r_distance
3. *A*[*i*] ←*ReadChptRec*(*A*[*i-1*].previous). /* reading of the previous checkpoint records */
4. **endfor**
5. *p_num* ← number of AP's saved in *R*.ap[].
6. *needed_local*[1,..,*R*.distance] ← 0. /* number of local checkpoints created for making a GCS */
7. *costs*[1,..,*R*.distance] ← 0. /* initialization */
8. **for** *i* = 1 **to** *R*.r_distance
9. **for** *j* = **to** *p_num*
10. **if** (**A**[*i*].dep_vec[j] > *R*.serial[j])
11. *needed_local*[[*j*]++.
12. **endif**
13. **endfor**
14. *costs*[*i*] = *i* + *needed_local*[*i*].
15. **endfor**
16. Find the least element among *costs*[1], *costs*[2],.., *costs*[*R*.distance] and let *s* be the index of that element.
17. **forall** AP p such that *A*[*s*].dep_vec[p] > *R*.serial[p]
18. Send a checkpoint requesting message to the logging agent managing the checkpoint record of *p*.
19. **endfor**
20. Blocked until all the response messaged are received form the logging agent's above.

Fig. 5. Algorithm for routine *CreateGCS*().

sent by Pi. If that is true, the routine saves the content of m in the checkpoint record being located in memory. Otherwise, if m is an inbound message, that is, m is a message coming from other logging agent, then the data of checkpoint's dependency and the serial numbers of the local checkpoints of other AP's are updated to reflect the changes of in the distributed application state.

4. Performance Analysis

To analyze the performance of the proposed scheme, we consider two key metrics, that is, less overhead paid for making checkpoints during normal execution time and low possibility of rollbacked executions in the face of application failure. Because there is a trade-off between these two metrics and they are largely affected by diverse factors such as application or network failure rates and frequency of checkpoint creations, it is very hard to devise an exact performance metrics. For these limitations, we give only a rough analysis on our checkpointing scheme here. ,

We first look on the overhead paid to generate checkpoint records during normal time. The main components of such overhead cost seem to be AP's blocking time for checkpointing and communication costs for sending

Algorithm: Routine *UpdateChptRec(m, R)*
INPUT: received message *m*, *Pi's* checkpoint record *R*

1. **if**(*m* is an outbound message) /* heading for other AP */
2. Save *m* into *R* and advance the number of sent messages by one.
3. **else** /* *M* is a message arriving at *Pi* */
4. **foreach** p in *R*.ap[] /* participant application processes */
5. *R*.serial[p] ← max(*m*.serial[p], *R*.serial[p]).
6. *R*.dep_vec[p] ← max(*m*.dep_vec[p], *R*.dep_vec[p]).
7. **endfor**
8. **endif**

Fig. 6. Algorithm for routine *UpdateChptRec*().

additional data used for tracking execution states of ongoing application. The blocking time in our algorithm is very short on average, compared with the traditional algorithms. To see that, recall the steps of lines 5-7 in Fig . 3. In those steps, blocking time arises only when a globally consistent local checkpoint is not found within the R-distance. In many cases, such a situation is not the case. Even though there is a need for creating a new global checkpoint, our protocol will choose a local checkpoint whose checkpointing overhead is most cheap. That is done by the logging agent by using the routine *CreateGCS*() of Fig. 5. Using this routine, the logging agent can choose among previous local checkpoints any one that demands a least number of local checkpoints.

The network cost for checkpointing depends on the amount of additional checkpoint-related data that piggybacks on messages delivering application data. In particular, such data should be less on the warless communications. To reduce the additional data on wireless networks, the logging agent manages checkpoint-related information using its checkpoint record in memory and refers to that for generating messages being transferred among logging agent's. From this, additional network overheads for checkpointing are small in our scheme, because most of additional data are not visible to AP's,

In the aspect of less cancellation of execution results in the case of application failure, our algorithm has a good property. Due to R-distance, the number of rollbacks in a particular AP is always less than a value set to R-distance. To all AP's participating in the application, the worst case number of rollbacks is less than $N \times (d\text{-}1)$ while N AP's are running with R-distance of d. That is, there is a tight upper bound on the number of cancelled local checkpoints. However, such a large cancellation is not realistic, since the times of checkpoint creations are different among the AP's joining a distributed transaction. In probabilistic, the

average number of rollbacked local checkpoints remains below a half of the upper bound number, that is, $N \times (d\text{-}1)/2$. Therefore, by setting R-distance appropriately, we can give a limit on the losses of execution results. From these properties, we can say that our algorithm can have a less cancellation of application by setting R-distance in a low range. Consequently, the proposed method has advantages during the normal execution time and recovery phase.

5. Conclusions

The problem of making a consistent distributed checkpoint in a mobile network environment is challenging to solve. This is because the distributed application needs application processes' communication via warless networks and such wireless communications easily make the cost for checkpoint higher. To have less communication cost, the previous checkpoint schemes for mobile distributed applications take an approach to making local checkpoints in a very inflexible manner. Such inflexible in the checkpointing causes the lack of semantics-awareness in the time of checkpointing. In addition, some asynchronized checkpointing scheme cannot provide any efficient mechanism for global checkpointing by the application process. These shortcomings can make obsolete checkpoints and frequent losses of expensive execution results. To solve those problems, we proposed a new checkpoint scheme based on the checkpoint agent and the concept of the recovery distance. From the combination of them, the proposed scheme provides the capability of semantics-aware checkpointing by paying only a cheap cost. We believe that the proposed checkpointing scheme can be applied to recover the mobile distribute application from diverse failures.

References

[1] T. Imielinski and B. R. Badrinath, Mobile Wireless Computing: Challenges in Data Management,Communications of the ACM, pp.19-28, Vol.37, No.10, October 1994.

[2] Yi-Bing Lin, Failure Restoration of Mobility Databases for Personal Communication Networks, Wireless Networks, Vol.1, No.3, 1995.

[3] Sashidhar Gadiraju and Vijay Kumar, Recovery in the Mobile Wireless Environment Using Mobile Agents,IEEE Trans. on Mobile Computing, Vol.3, No.2, April 2004.

[4] Ricardo Baratto, Shaya Potter, Gong Su, and Jason Nieh, MobiDesk: Mobile Virtual Desktop Computing, In Proc of the 10th International Conference on Mobile Computing and Networking, pp.1-15, 2004.

[5] Dhiraj K. Pradhan, P. Krishna, and Nitin H. Vaidya,Recovery in Mobile Wireless Environment: Design andTrade-off Analysis, In Proc. of the 26th International Symposium on Fault-Tolerant Computing, pp.16-25,1996.

[6] Arup Acharya and B. R. Badrinath, Checkpointing Distributed Applications on Mobile Computers, In Proc. of the 3rd International Conference on Parallel and Distributed Information Systems, pp.73-80, 1994.

[7] Y. M. Wang, Consistent Gobal Chckpoints That Contain a Given Set of Local Checkpoints, IEEE Trans. on Computers, Vol.46, No.4, pp.456-468, 1997.

[8] Tongchit Tantikul and D. Manivannan, Communication-Induced Checkpointing and Asynchronous Recovery Protocol for Mobile Computing Systems, In Proc. of the 6th International Conference on Parallel and Distributed Computing Applications and Technologies, pp.70-74, 2005.

[9] Taesoon Park and Heon Y. Yeom, An Asynchronous Recovery Scheme based on Optimistic Message Logging for Mobile Computing Systems, In Proc. of the 20th International Conference on Distributed Computing Systems, pp.436-443. 2000.

[10] R. E. Strong and S. Yemini, Optimitic Recovery in Distributed Systems, ACM Trans. on Computer Systems, Vol.3, No.3, August 1985.

[11] D. Manivannan and Mukesh Singhal, Quasi-Synchronous Checkpointing: Models, Characterization, and Classification, IEEE Trans. on Parallel and Distributed Systems, Vol.10, No.7, July 1999.

[12] Cheng-Min Lin and Chyi-Ren Dow, Efficient Checkpoint-based Failure Recovery Techniques in Mobile Computing Systems, Journal of Information Science and Engineering, pp.549-573, Vol 17, No.4, 2001.

[13] Lorenzo Alvisi, E. N. Elnozahy, Sriram Rao, Syed Amir Husain and Asanka De Mel, An Analysis of Communication Induced Checkpointing, In Proc. of the Symposium on Fault-Tolerant Computing Symp., pp.242-249, 1999.

[14] Franco Zambonelli, On the Effectiveness of Distributed Checkpoint Algorithms for Domino-Free Recovery, In Proc. of High Performance Distributed Computing,pp.124-131, 1998.

[15] Mootaz Elnozahy, et. al., A Survey of Rollback-Recovery Protocols in Message-Passing Systems, Technical Report: CMU-CS-99-148, June 1999.

[16] Yi-Min Wang and W. Kent Fuchs, Lazy Checkpointing Coordination for Bounding Rollback Propagation, In Proc. of the International Symposium on Reliable Distributed Systems, pp.78-85, 1993.

[17] Lapmport, Time, clocks, and the Ordering of Events in a Distributed System, Communication of ACM, Von. 21, No.7, pp.558-565, 1978.

Sungchae Lim received the B.S. degree in Computer Engineering form Seoul National University at 1992, and achieved the M.S. and Ph.D. degrees in Computer Science from Korea Advanced Institute of Science and Technology (KAIST), at 1994 and 2003, respectively. He also worked for the Korea Wisenut Cooperation from 2000 to 2005, and he is currently an Associate Professor in the Department of Computer Science at Dongduk Women's University. His research interest includes the high-performance indexing, mobile computing, and semantic Web.

Implementation of MDA Method into SOA Environment for Enterprise Integration

Wiranto Herry Utomo

Faculty of Information Technology, Satya Wacana Christian University
Salatiga, Central Java, Indonesia

Abstract

Even though SOA provides real contribution, it is not adequate to implement enterprise integration. There are still problems in the implementation of enterprise integration in SOA environment, they are 1) the absence of modeling language support, 2) the absence of guideline of the services implementation produced by services identification, and 3) service orchestration that uses only Web Services. Based on the consideration and comparison of some integration methods, MDA method from OMG is chosen as a method to help dealing with SOA weaknesses. Model-driven based MDA method enables business level functionality to be modeled by UML language modeling that is separated from low level implementation (code level). Therefore, SOA used in MDA approach can be expressed using UML modeling language. This study proves that SOA-MDA method has been successfully used to perform analysis, design and implement of enterprise integration.

Keywords: *SOA, MDA, UML, Web Services, Integration.*

1. Introduction

SOA is a framework in company architecture and aims at achieving the same business' goals: minimize ownership costs and create flexible business solutions that improve business' stability, reduce time to the market and provide support for global expansion. SOA substantially impacts the whole key aspects of enterprise architecture. Business service proposed by SOA forms the basic of business architecture and process architecture.

SOA forms business architecture because business' functions are exposed as services that can be divided and reused. Business process, services and event are converted to appropriate application services that create and support services architecture. Services alone form application architecture, whereas information architecture is achieved through data standardization and data availability through interface services [17]

SOA is a software architecture designed based on service oriented design principles [3][15][5][8][13][7], whereas service orientation is a concept in software engineering that represents different approaches to separate interest.

According to Erl [3], in general, software that does not use SOA can be divided into two main layers, Application Layer where application runs and Business Process Layer that describes how business process in a company runs. Organization business process will be defined in application along with technical program code. In SOA implementation, service oriented process is implemented in a layer between Business Process and Application Layer where both are parts of logic enterprise. The layer is called Service Interface Layer and can be seen in Figure 1.

This layer is to wrap the logic in Application Logic along with the business process in Business Logic. Through this approach, application can be more modularized with more varied technology.

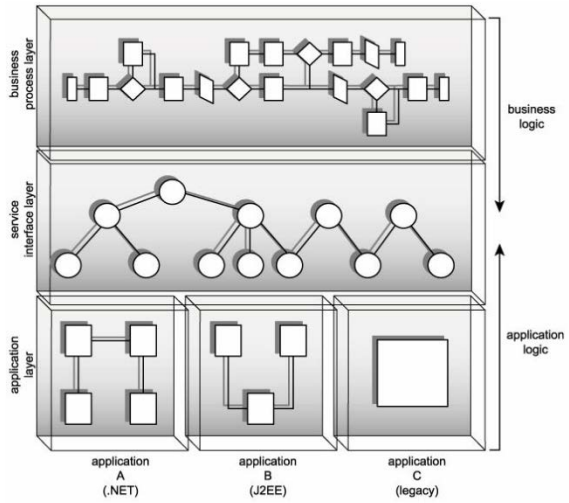


Fig 1. Service Interface Layer in SOA method [3]

Services analyses and identification can be done through this SOA method [3]. The services achieved then mapped to Service Interface Layer, which are Application Layer, Business Process Layer and Service Interface Layer (Figure 1).

Even though SOA [3] provides real contribution, it is not adequate to implement enterprise integration. There are still problems in the implementation of enterprise integration in SOA environment, they are 1) the absence of modeling language support, 2) the absence of guideline of the services implementation produced by services identification, and 3) service orchestration that uses only Web Services.

Orchestration using Web Services has two weaknesses, in terms of scalability and the inability to deal with protocol and data discrepancy. To deal with this, Web Services orchestration with ESB has now been developed. ESB is an infrastructure for SOA service connection and message exchange. ESB main functionality is to do routing, protocol transformation and message or data transformation. Protocol and data discrepancy can be overcome by the protocol and data transformation in ESB. ESB eases connection and mediation, simplifies integration and eases the reuse of service components that lead to a high scalability integration.

Other strengths of Web Services orchestration with ESB is that it enables business layer and information system to have a closer relation because Web Services orchestration is presented in high abstraction level called business process by hiding middleware traditional object used to support business to business interaction. Aside from that, business requirements can be directly translated into business process application through Web Services composition. That is why SOA method alone is not yet optimal to implement enterprise integration. Thus, other methods that are able to deal with the method's weaknesses are required.

Based on the consideration and comparison of some integration methods, MDA method from OMG is chosen as a method to help dealing with SOA weaknesses. The decision to choose MDA method to be combined with SOA method is based on: 1) MDA method is a model-driven method based on the use of platform independent technology model, 2) this method can be used to transform high level business process model to low level one (code), 3) the existence of standard modeling language, 4) this method has used ESB as midleware infrastructure, 5) the phases of the process in this method use system development life cycle.

Model-driven based MDA method [9][4] enables business level functionality to be modeled by UML language modeling that is separated from low level implementation (code level). Therefore, SOA used in MDA approach can be expressed using UML modeling language.

By combining MDA and SOA methods, two completing each other advantages will be gained. SOA provides an infrastructure that reduces complexity in the services reuse and integrates all kinds of technology, protocol, and application whereas MDA is used in High Level Business Process Model transformation to platform independent low level one (programming code). The integration of SOA and MDA methods will be a complete method for enterprise integration..

2. Related Works

Rafe et al [12] stated that the goal of their research are to provide a successful and usable conjunction between these two technologies. They have tried to provide a simple yet effective process which can be viewed as a framework. In the vision inspired by this framework, SOA is the product and MDA makes its production line. During this process, input model is provided via XMI standard and with a high level of abstraction. Proposed framework analyses the elements and their relations within the given model and tries to recognize the SOA components. In two phases (Figure 2), the input model is first transformed into a SOA profile based model and then into a middleware independent code. Middleware transparency is achieved via the concept of Aspect. The final phase of framework is to transform middleware transparent code into an executable code based on one of known middlewares for SOA. Jini middleware and pre-process weaving is used in the last phase.

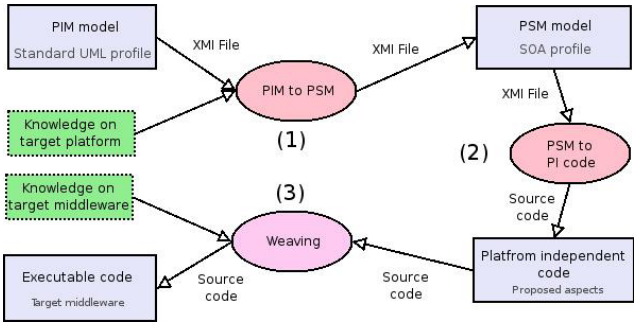


Fig 2. Framework Components [12]

The phase of PIM to PSM can be considered as the most important and complex part of the framework. In this step, the platform independent model - based on UML standard profile - is transformed to the platform specific model - based on proposed SOA profile. Although we have tried to apply MDA to SOA for simpler model, the approach taken here has more capabilities and can handle relatively more complex cases.

In this approach, the input model (PIM) has no direct information about SOA. Obviously using such an abstract input - based on standard UML - requirements a more autonomous model transformer. By autonomous we mean

a model transformer which tried to depend on the specification of model rather than human guidelines. Such a model transformation is beyond what we expect from an MDA based model transformer and also beyond most of the current frameworks.

Model-Driven Architecture (MDA) is proposed by the Object Management Group (OMG) as a reference to achieve wide integration of enterprise models and software applications. MDA is a best choice to address how SOA should be designed, developed and integrated. MDA provides specifications for an open architecture appropriate for the integration of systems at different levels of abstraction and through the entire information systems' life-cycle. The MDA comprises three main layers: Computation-Independent Model (CIM), Platform-Independent Model (PIM), Platform Specific Model (PSM) . MDA lies in separating the enterprise model from the technology infrastructure, making a clear division between the business functions and the implementation details.

The Computation Independent Model (CIM) cares about the requirements for the systems by describing the situation in which the system will be used. Such a model is sometimes called a domain model or a business model and hides information about the use of automated data processing systems.

The Platform-Independent Model (PIM) describes the operation of a system while hiding the details necessary for a particular platform. The model focus on specifications that are not changing from one platform to another e.g. BPMN (independent from Workflow engine) or UML (independent of computing platform).

A Platform-Specific Model (PSM) combines the specifications in the PIM with the details that specify how theses systems are using a specific type of platform.

There are three levels, CIM, PIM and PSM, in MDA method according to OMG. In the research conducted by Rafe et al [12], CIM level is not used but is directly jumped to PIM-PSM level. Aside from that, in Rafe et al research [12], SOA is used after modeling of the PIM-PSM level. Therefore, therefore two differences of this research and Rafe et al research [12] : 1) this research used three complete MDA levels, CIM, PIM and PSM, 2) SOA is combined in CIM level to MDA in order to decide business process and identify services, MDA modeling in the next level, PIM and PSM, is done after the services found.

3. SOA – MDA Methods

The integration OF SOA and MDA complete each other and cover each other weaknesses. Actually, the use of 'integration' term is not appropriate. The more appropriate term is MDA 'implementation' into SOA environment. The application of model driven method in SOA environment is phases of high level Business Process model into executable services and can be orchestrated into the integration of services. Phases or processes of the integration method refer to Object Oriented System Development Life Cycle that refers to Solamo [14]. Phases of SOA-MDA method can be seen in Figure 3.

New method as a result of integration proposes service oriented approach for integration by determining two factors:

- Business Perspective focuses on business features and requirements from the application that will be constructed.
- System Perspective focuses on functionality and process requirements to be implemented in the application to satisfy business requirements.

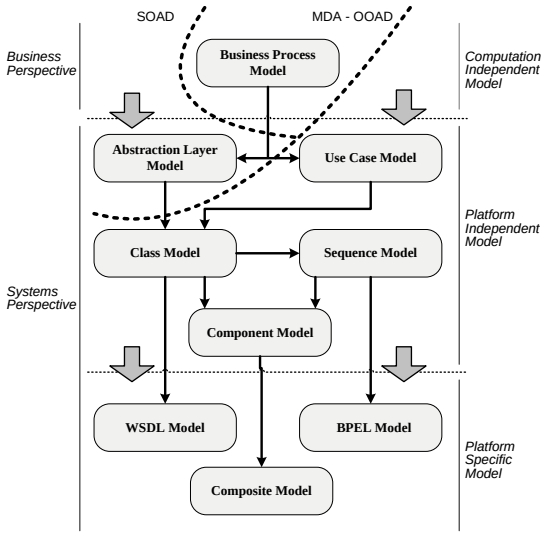


Fig. 3 New method of the integration of SOA and MDA method

This integration method provides a series of concepts required for modeling of the two perspectives. All concepts can be seen in Figure 3 that represents the two methods, SOA and MDA. The concepts related to business perspective explain the attached elements in business and are represented in CIM Model through Business Process Model. The concepts related to system perspective are elements used to describe system functionality and process and are represented in PIM dan PSM Model, with Use Case Model, Class Model, Sequence Model, Component Model, WSDL Model, BPEL Model and Composite Model.

Some barriers in the selection of programming language, hardware, network typology, communication protocol, infrastructure, etc occur in software development. Each element is considered as a part of the platform solution. CIM approach helps to focus on essential part of the solution designed, separated from platform details. CIM does not show structure details of the system. CIM plays important roles in bridging the gap between domain expert and its requirements, as well as the experts who constructs artifacts who work side by side to fulfill the domain requirements. CIM is referred as business domain model that explains the knowledge of business domain, which is free from business process or particular software used [1][16][6][10].

In CIM level, which is business analysts oriented, this method uses Business Process Model by adopting BPMN notation that is in fact the standard of business process modeling. However, the business process modeling also uses Activity diagram beside BPMN notation. Business Process Model is used to define identification guideline and business concept representation. This Business Process Model eases service identification to be implemented into application, and transform it into low level one such as PSM level to model Web Services and its composition.

PIM is a view of a system from platform independent point of view. PIM indicates particular levels of platform-freedom so that it can be used for some different platforms. PIM can be seen as the specification of free technology system functionality that will be used to implement the functionality. PIM provides formal specification from system structure and function that is free from any platform. From this point of view, it can be said that CIM is a component of PIM since platform independent component describes computational component and its independent interaction. This components and interface are ways to realize some more abstract information system or application, which automatically help to create a CIM [1][16][6][10].

PSM explains how particular technology can be used to implement the function described in PIM. PSM is adapted with the system in term of implementation construction provided by a particular implementation technology. PSM possesses components for target platform. PIM can be transformed into one or more PSM. Particular platform is produced for every particular technology platform [1][16][6][10].

According to OMG [9][4], PSM is a system reviewed from a particular platform point of view. For Example, Class Model is PIM with service implementation architecture choice, if the model chooses to use particular service technology such as Web Services, the Class Model is then transformed into specific PSM for Web Services.

The use of UML can be said as a common thing in most methods. However, this integration method proposes the use of UML for modeling notation in PIM level, whereas modeling in PSM level will be adapted with implementation platform.

This integration method is a complete development method because it views modeling from all elements related to services oriented system development. This method does not only include one level in driven model, such as PIM, or PSM, but it includes all level including CIM, PIM dan PSM.

4. PHASES OF SOA-MDA METHOD

Figure 3 shows that the processes start by building Business Process Model that later result in services. These processes include several phases where each phase is related to the “creation” of different models. The phases of this integration model include nine phases 1) Business Process Model, 2) Abstraction Layer Model, 3) Use Case Model, 4) Class Model, 5) Sequence Model, 6) Component Model, 7) WSDL Model, 8) BPEL Model and 9) Composite Model. The phases of development process of this integration method are:

1. Business Process Model

With the fast changing business, company can build new business processes by running existing application. This model includes in business perspective and CIM layer. This model is a high level model that serves as the modeling starting point of this method. This model is derived from SOA and MDA methods. The notation used in Business Process Model can use either BPMN or Activity Diagram UML notation.

Business Process Model does not only focus in individual business process representation, which can be fulfilled by workflow description using BPMN that are implemented into WS-BPEL, but also focuses in the development of SOA solution using business process. Business Process Model manages services in workflow business context. This model displays service management in top-down process level. Top-down direction eases the mapping of business requirements into tasks that include activity flows, every is activity realized by existing business process and service components.

To decompose business process, first, tasks are broken down into smaller ones and then map each business process into services.

By implementing SOA design and method, company business processes are modeled and processes blocks that can be grouped into services are identified. Legacy application is analyzed based on its functionality and is mapped to the services. New service is constructed when there is business process that cannot be mapped to legacy application.

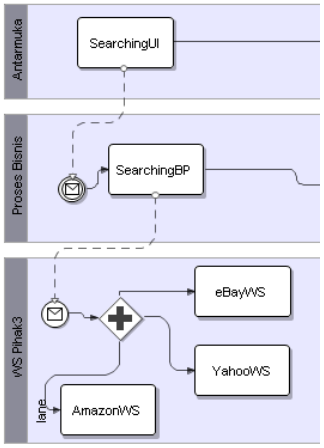


Fig. 4 Process Model using BPMN notation

2. Abstraction Layer Model

Services decomposition can be done from Business Process Model by decomposing business process into the smaller ones. Process Model can be generated into services required to construct new business process. These services can be developed from legacy application, third party or constructing new services.

Abstraction Layer improves Web Services group concept that is a group of Web Services that serves business function as general. Web Services can be published by different service provider and be differentiated from others through specific features. Service layer shows top-down or bottom-up service layer handling.

Even though this integration method is the integration of SOA methods from Erl [3], there are some differences in this Service Layer Model. These differences occur because this integration method uses middleware ESB via BPEL to do Web Services orchestration, whereas in SOA method, Web Services orchestration still uses Web Services (in Service Layer Orchestration).

Service candidate identification is carried out in every layer in SOA, which exists in Application Service Layer, Business Service Layer, and Orchestration Service Layer. However, as mentioned before, ESB replaces Orchestration Service Layer. Therefore, services candidate identification is only carried out in two SOA layers, Application Service Layer and Business Service Layer. Service candidate identification is carried out based on the requirements derived from application Use Case Model and Business Process Model.

In Business Service Layer, identification is carried out by looking at the existing business process and the parts nominated as service candidates are identified. The identification process of service candidate in this layer can be done through task-centric business service. In identifying task-centric business service, identification from Use Case Model in Use Case Diagram is also done in addition to the use of existing business process. Task-centric business layer is gained by mapping the existing phases in business process into services.

Based on Business Process Model constructed, two kinds of services in this layer can be seen 1) input services that create trigger toward business process, and 2) output services that promote invoke. Rademakers-Dirksen [11] explicitly divides these services into two, inbound service to carry out input connection configuration and outbound to carry out output connection configuration. Referring to Rademakers-Dirksen [11], there are two kinds of services in this layer, inbound service and outbound service.

Therefore, in this integration method, Abstraction Layer Model is the improvement of Erl layer model (2005), with the following improvement:

1. Orchestration Service Layer is improved into Service Bus Layer
2. Leave out Application Service Layer. This layer is left out because by the implementation of ESB for integration, all services are services related to business process, and there is no services that technically related to only Application Layer.
3. Business Service Layer is grouped into two service layers, Inbound Service Layer and Outbound Service Layer.

By the use of this new Abstraction Layer Model, the former services found are mapped into Abstraction Layer Model as shown in Figure 5.

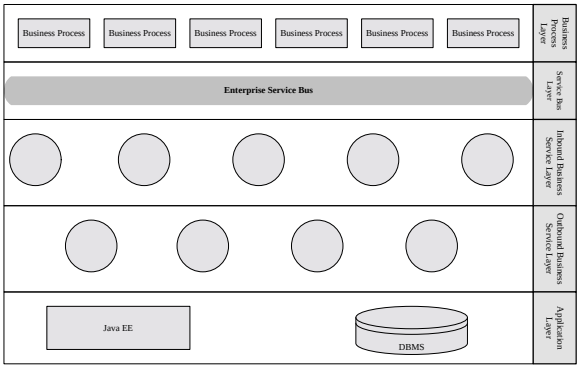


Fig. 5 Abstraction Layer Model

3. Use Case Model

This Use Case Model describes the system's requirements. This phase is a form of software engineering that enables developer in understanding problem domain. This Requirement Model is a series of tasks to know the impact of software development, what the costumers want, and how end users will interact with software.

Use Case Model is used to describe what the system will do, system's functional requirements, and the expected system functionality along with its environment. Complement specification is the not-yet-mapped requirement into Use Case specification that includes non functional requirements such as code maintenance, performance reliability, and system supports or obstacles as well as safety. Use Case Model is a mechanism to achieve expected system behavior without determining how behavior system is implemented.

The output produced in this phase is in the form of Use Case Model (Use Case Diagram), and Use Case Specification. This Use Case model is usually gained based on user's requirements, however, in this integration method, Use Case is produced from Business Process Model. Every business process, which is a functionality of a business unit, is represented into every Use Case of Use Case Diagram.

Use Case Diagram consists of Actor and Use Cases. This diagram shows system functionality and actor communicate with the system. Every Use Case in the model explains the details of the use of Use Case specification. Use Case UML diagram is used as modeling tool for Use Case model. This Use Case diagram consists of three components (see Figure 7):

1. Actor, represents a series of role played by users or system when they interact with Use Case. The actor calls the system to send a service. This Actor can be human or other system. Actor is named after nouns.
2. Use Case, describes the function displayed by the system when it interacts with the Actor. This Use Case is described using verbs or verb phrases.
3. Association, shows the relation or association between Actor and Use Case and or inter-Use Case

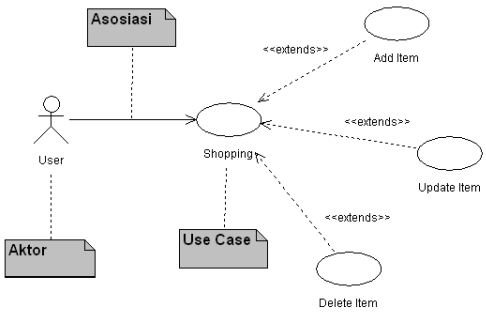


Fig. 6 The Example of Use Case Diagram

4. Class Model

Class model is constructed using UML Class Diagram. This Class Diagram is an input for the following program development. This Class Model represents previous conceptual model over something in the system that possesses behavior. Class Model is an important model in software development because it will be the main input for the following phase. This model is described in Class Diagram containing classes that provide former conceptual model for things in the system that possesses property and behavior. This Class Diagram consists of Boundary Class, Control Class, Entity Class and Web Services Class.

There are four perspectives used in identifying classes, they are boundary between system and actor, the information used by system, and control logic of the system. These four perspectives are described into classes, they are: Boundary Class, Control Class, Entity Class and Web Services Class (Figure 7).

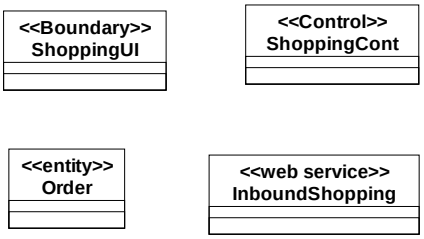


Fig. 7 The Example of Boundary Class, Control Class, Entity Class and Web Services Class

Boundary Class is used to model the interaction between environment and the system working in it. This Class explains system boundary and starting point in identifying related services. This Class limits external power from internal mechanism and vice versa. This Class bridges interface and something outside the system that consists of interface users, system interface and interface tools.

Boundary Class is derived from Use Case Diagram, originally from the set of Actor and Use Case.

Control Class represents system functionality. This Class provides behavior that defines control logic and transaction in Use Case, contributes small change if Entity Class' structure or behavior changes, uses or governs some Entity class' contents. This Class provides system coordinated behavior and limits Boundary Class and Entity Class.

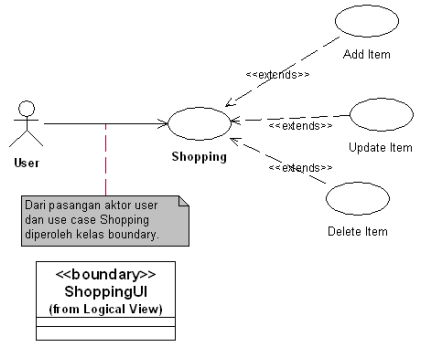


Fig. 8 Boundary Class derived from the set of actor and Use Case

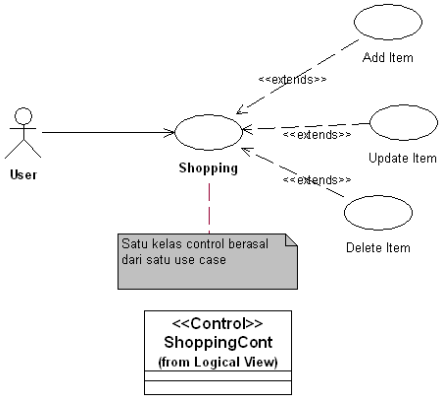


Fig. 9. Control Class from Use Case

Entity Class represents information storage in the system. This Class is used to update information, such as events, phenomena or objects. This class responsible for storing and managing information in the system that represents key concept from the system constructed. This Class Entity can be derived from key abstraction of Use Case Diagram by filtering noun.

Web Services Class is a representation of services found in Service Layer Model.

5. Sequence Model

Sequence Model is created using UML Interaction Diagram containing Collaboration Diagram and Sequence Diagram. Interaction diagram models dynamic characteristics of objects in groups of classes. This diagram models system behavior as how system responds toward a particular user's action, how an object is created or changed as well as how data is transformed.

This model shows interaction and collaboration among analyses classes. Two basic elements used in Behavior Model are object and message. Object is an instantiation of a class, whereas message is a form of communication among objects. This service Process Model can be seen as a collaboration among the object of classes in Service Model. Therefore, the input from this Sequence Model is derived from Class Model. An example of Sequence Model can be seen in Figure 10.

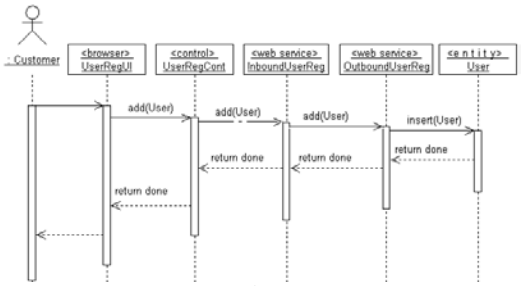


Fig. 10 An Example of Service Process Model

6. Component Model

This model expands the representation of Class Model and the Sequence Model modeled before. Component Model represents components from Web Services composition that identifies services collaborating with business process. This method represents this model using Component Diagram.

This Component Diagram describes components integration which in the next phase will be derived into composite application. The example of Component Model can be seen in Figure 11.

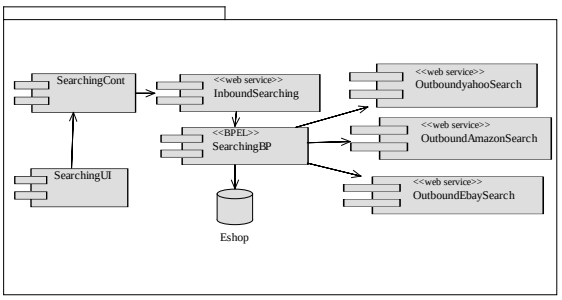


Fig. 11 An Example of Component Model

7. WSDL Model

This model is used to describe Web Services interface that will be used to deliver every services provided by the system. This model is based on WSDL standard. WSDL is a language proposed by W3C to describe Web Services and enables it to describe the interface of services in XML

format. WSDL Model allows to derive graphic representation of Web Services interface that will be generated into WSDL code automatically. The example of Web Service Interface Model implemented in Java EE platform using Netbeans.

8. BPEL Model

This model expands service composition identified by process model explained above by adding particular Web Services based platform details. This Process Execution Model is represented in the form of WS-BPEL. WS-BPEL is a Web Services extension used to facilitate modeling process and BPEL execution in Web Services. BPEL is a modeling language in XML format used to describe business process. The model produced by this language is later executed by BPEL engine.

The explanation of elements in this language will be explained as follow [3]:

1. Process. Process is BPEL's main element. The name of process is defined as name attribute. Aside from that, this tag is also used to insert information related to process definition.
2. PartnerLink and partnerLinks. This element defines the kinds of port from other services involved in business process execution.
3. Variables. This element is used to keep status information used during the process of workflow logic.
4. Sequence. This element organizes a group of activities that they can be executed in an orderly manner. Whereas the elements are supported by WS-BPEL for sequence such as recieve, assign, invoke, and reply.

Beside the four main elements above, WS-BPEL also facilitates some other tags. Standard from BPEL is defined by OASIS and can be achieved from OASIS website. The example of BPEL Model implemented in BPEL using Netbeans can be seen in Figure 12.

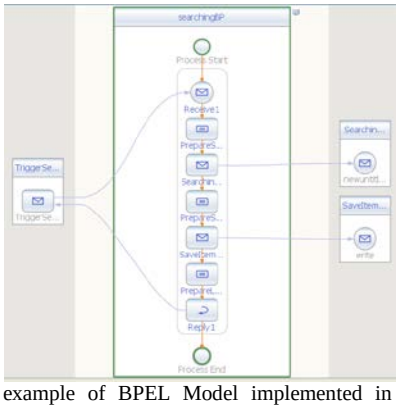


Fig. 12 An example of BPEL Model implemented in BPEL using Netbeans.

9. Composite Model

Composite application (SOA composite application) according to Binildas [2] is an SOA application containing some components such as services, BPEL process, ESB mediation, rules, adapter, and etc. All components must cooperate and support one or more Composite Application.

This model is SOA application modeling containing some components such as BPEL process and ESB. All the components cooperate and support one or more Composite Application. Composite application is the integration of services containing business function and information from a separate source of information. Composite application is a form of integration and application development. Specifically, Composite Application is constructed to support company business process and map it to underlying information resources. In business integration, Composite Application is the final product of SOA. The example of Composite Application can be seen in Figure 13.

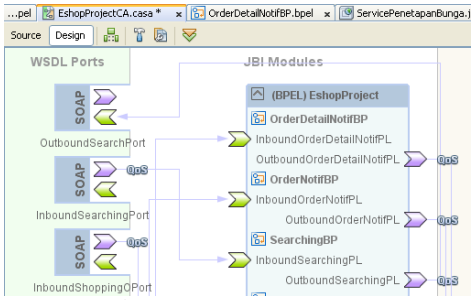


Fig. 13 An example of Composite Application implemented in CASA using Netbeans.

4. Conclusions

The case study to proves thisn methods consists of e Shop application where consumers can shop and place orders for goods offered for sale there. The e-Shop doesn't store inventory but it relies on third parties to warehouse and ship the goods. The third party consisted of the Amazon, Ebay and Paypal. As soon as the e-shop receives an order, it creates a purchase order and sends it to the backend purchasing system which, in turn, sends orders out to one or more suppliers for fulfillment.

This case study of e Shop application proves that SOA-MDA method has been successfully used to perform analysis, design and implement of enterprise integration.

References

[1] Almeida, J.P.A., 2006, Model-Driven Design of Distributed Applications, Ph.D. Thesis, Centre for Telematics and Information Technology, University of Twente, Netherlands

[2] Binildas, C. A., 2008. Service Oriented Java Business Integration, Birmingham-Mumbai: Packt Publishing.
[3] Erl, T., 2005, Service-Oriented Architecture: Concepts, Technology, and Design, Prentice Hall PTR, Upper Saddle River, New Jersey 07458
[4] Frankel, D.S., 2003, Model Driven Architecture : Applying MDA to Enterprise Computing, Wiley Publishing, Inc., Indianapolis, Indiana
[5] Kanchanavipu, K., 2008, An Integrated Model for SOA Governance An Enterprise Perspective, Master Thesis, IT University of Göteborg Chalmers University of Technology and University of Gothenburg, Göteborg, Sweden
[6] Kim, H., 2008, Modeling of Distributed Systems with SOA & MDA, IAENG International Journal of Cumputer Science, 35:4, 20 November 2008
[7] Li, G., Muthusamy,V. and Jacobsen, H., 2010, A Distributed Service-Oriented Architecture for Business Process Execution, ACM Transactions on The Web, Vol. 4, No. 1, Article 2, Publication date: January 2010.
[8] Nikayin, F.A., 2009, Adopting A Theoretical Method For The Development Of A Service-Oriented Information System, Dissertation, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur
[9] Pastor, O. and Molina, J.C., 2007, Model-Driven Architecture in Practice A Software Production Environment Based on Conceptual Modeling, Springer-Verlag Berlin Heidelberg
[10] Pokraev, S.V., 2009, Model-Driven Semantic Integration of Service-Oriented Applications, Ph.D. Thesis, Centre for Telematics and Information Technology, University of Twente, Netherlands
[11] Rademakers, T., dan Dirksen, J., 2009, Open Source ESBs in Action, Manning Publications Co., Greenwich, CT 06830
[12] Rafe, V., Rafeh, R., Fakhri, P., and Zangaraki, S., 2009, Using MDA for Developing SOA-Based Applications, International Conference on Computr Technology and Development, IEEE Computer Society
[13] Reddy, V.K., Dubey, A., Lakshmanan, S., Sukumaran, S. and Sisodia, R., 2009, Evaluating legacy assets in the context of migration to SOA, Software Qual Journal (2009) 17:51–63, Springer Science+Business Media
[14] Solamo, R., Antonio, J., Asrani, N., Chen, D., de Guzman, O., Feria, R., Petines, J.P., Shin, S., Srinivas, R., Thompson, M. and Villafuerte, D., 2006, Software Engineering, Java Education & Development Initiative, Sun Microsystem.
[15] Sterff, A., 2006, Analysis of Service-Oriented Architectures from a business and an IT perspective, Master Thesis, Technische Universität München, Fakultät für Informatik
[16] Vidales, M.A.S., García1, A.M.F., and Aguilar, L.J., 2008, A new MDA approach based on BPM and SOA to improve software development process, Tékhne, 2008, Vol VI, no 9, ISSN: 1645-9911
[17] Vos, W., and Matthee, M.C., 2011, Towards A Service-Oriented Architecture: A Framework For The Design Of Financial Trading Applications In The South African Investment Banking Environment, South African Journal of Industrial Engineering May 2011 Vol 22(1)

Dr. Wiranto Herry Utomo is an associate professor of the Departement of Information System at the Satya Wacana Christian University, Salatiga, Central Java, Indonesia. He has published in several journals including IJWA, MASAUM Journals, and international conference including iiWAS of the ACM. His research interests include the enterprise integration, strategic alignment, SOA, Web Services, BPEL, Enterprise Service Bus, and Java EE.

Comparison and Application of Metaheuristic Population-Based Optimization Algorithms in Manufacturing Automation

Rhythm Suren Wadhwa[1], Zhenyou Zhang[2] , Quan Yu[2] and Kesheng Wang[2]

[1,2] Inst. for produksjons- og kvalitetstek., NTNU
Trondheim, 7491, Norway

Abstract

The paper presents a comparison and application of metaheuristic population-based optimization algorithms to a flexible manufacturing automation scenario in a metacasting foundry. It presents a novel application and comparison of Bee Colony Algorithm (BCA) with variations of Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) for object recognition problem in a robot material handling system. To enable robust pick and place activity of metalcasted parts by a six axis industrial robot manipulator, it is important that the correct orientation of the parts is input to the manipulator, via the digital image captured by the vision system. This information is then used for orienting the robot gripper to grip the part from a moving conveyor belt. The objective is to find the reference templates on the manufactured parts from the target landscape picture which may contain noise. The Normalized cross-correlation (NCC) function is used as an objection function in the optimization procedure. The ultimate goal is to test improved algorithms that could prove useful in practical manufacturing automation scenarios.

Keywords: *Bee Colony Algorithm, Particle Swarm Optimization, Ant Colony Optimization, Foundry Automation*

1. Introduction

In the 21st century, under the influences of globalization, manufacturing companies are required to meet continuously changing customer demands. Flexible manufacturing systems (FMS) has emerged as a science and industrial practice to bring about solutions for unpredictable and frequently changing market conditions [21]. Existing FMS implementations in manufacturing companies have demonstrated a number of benefits by helping lower production costs, increased factory floor utilization, reduced work-in-process, etc. However, there are a number of problems faced during the life cycle of an FMS, which could be classified into work flow design, production leveling, and control problems [21]. In particular, the production leveling is important owing to the dynamic nature of FMS such as flexible machines, tools and workflow. This work is primarily concerned with production leveling problem. Over the last decade, most research in FMS has been focused on scheduling of FMSs for single or multi objective problems. The present work, however, compares three evolutionary computation techniques Particle Swarm Optimization (PSO), Bee Colony Algorithm (BCA) and Ant Colony Optimization (ACO). The goal of the paper is not to declare one of the techniques as better than the other, but to test their applications after modification to suit the manufacturing scenario discussed, as well as their limitations. The case study is a small-to-medium batch manufacturing foundry and we intend to test the suitability of the algorithms for the purpose of lean workflow and reducing machine starvation in the manufacturing facility.

1.1 Earlier Research

1.1.1 Flexible Manufacturing Systems

During the last two decades much research has been done in this area. The heuristic algorithms developed include enumerative procedures, mathematical programming and approximation techniques, i.e., linear programming, integer programming, goal programming, dynamic programming, network analysis, branch and bound, genetic algorithm (GA), etc.

Shankar and Tzen [39] considered scheduling problems in a random FMS as composite independent tasks. Lee [25] presented a goal-programming model for multiple conflicting objectives in manufacturing. Toker et al. [45] proposed an approximation algorithm for ‘n’ job ‘m’ machine problem. Steeke and Soldverg [43] investigated various operating strategies on a caterpillar FMS by means of deterministic simulation with the number of completed assemblies on a performance criterion manufacturing problem associated with parallel identical machines throughout simulation. Chan and Pak [3] proposed two heuristic algorithms for solving the scheduling problem with the goal of minimizing total cost in a statictically loaded FMS. Shaw and Winston [40] addressed an artificial intelligence approach to the scheduling of FMS. Schultz and Merkens [38] compared the performance of an ES, a GA and priority rules for production systems. Further, a comprehensive survey on FMS was done by Chan et al. [3].

Many authors have been trying to emphasize the utilization of heuristics in flexible manufacturing automation. In this context, it has been proposed a comparative study on the application of evolutionary algorithms in a specific manufacturing environment i.e. metalcasting foundries.

1.1.2 Object Recognition in Flexible Manufacturing

The challenge of object recognition is to develop the ability to recognize objects even with significant variations in visual appearance. In recent years, a number of metaheuristic algorithms have been proposed. They have been applied to several real world combinatorial problems in manufacturing. For example, Silva, Lopes and Lima [41] as well as Perlin, Lopes and Centeno [36] presented two metaheuristic approaches, one based on compact Genetic Algorithm (CGA) and the other based on Particle Swarm Optimization (PSO). Results show that both methods can be efficiently applied to practical situations with reasonable computational costs.

Some other related works have been presented using variations of metaheurisitc algorithms. Tereshko and Loengarov [44] proposed a collective decision model considering a bee colony as a dynamical system where intelligent decision making arises from an enhanced level of communication among individuals. In their work, they discussed how the information exchange between individuals leads to globally intelligent selection of food sources in an unpredictable environment. Karaboga [19] proposed the Artificial Bee Colony (ABC) algorithm, based on the foraging behavior of real bees, and later compared its performance with other evolutionary and swarm intelligence based algorithms using a large set of numerical functions. Karaboga et al. [19] concluded that the ABC algorithm is a robust optimization algorithm that can be efficiently used in the optimization of multimodal and multi-variable problems. Another version of a bee swarm-based algorithm was proposed by Pham and Zaidi [37], named Bees Algorithm (BA), which can be used for both combinatorial and multi-parameter functional optimization.

More recently, Hackel and Dippold [13] developed an algorithm inspired in bee colony for the vehicle routing problem with time windows. According to Mishra [30], the algorithms mentioned before have an inherent probabilistic nature and thus may not always obtain best solutions with certainty. This paper uses the Matlab toolbox from Karaboga which minimizes or maximizes functions. We have adapted it in order to be able to take 4 templates and landscape image and be able to maximize the NCC value obtained by the equation 1, which is defined as “objective function” for maximization. Plotting commands have been added to the program to represent the matching between both images and so to be able to determine the accuracy of the program. Another command to calculate the time expended in each run has been added as well.

2. Problem Description

One important application of a robot vision system is to recognize whether or not a given part is a member of a particular class of parts. Currently, common examples of object recognition can be found in areas such as industry, engineering, medical diagnosis etc. Generally, recognition of objects in images using traditional search algorithms is computationally expensive. For many industrial applications, these algorithms should normally be executed in real-time. Hence, fast algorithms are essential at all stages of the recognition process in images. This fact suggests the use of fast algorithms based on metaheuristics. Recently, besides the traditional image processing techniques, several methodologies based on computational intelligence have been developed and applied to object recognition problem, so as to reduce computational cost and to improve efficiency. Amongst them, metaheurisitc population-based optimization algorithms, such as those from the Swarm Intelligence area, were successfully applied to the problems.

Recognizing orientation of objects is a challenging task due to constant changes in images in the real world. The most straightforward technique for part orientation recognition is called template matching [2]. Template matching is the process of determining the optimal matching between the same scenes taken at different times or under different conditions and the template known according to some similarity measure. [26]. In other words, the basic idea is to find a match of the pattern in some part of the landscape image. The most common way of finding the matching point between the landscape image and the template is by calculating the correlation function value which indicates the percentage of matching of both images for a specific matching point. The bigger this parameter is, the closer the two images will be.

Normalized Cross Correlation (NCC) is the most robust correlation measure for determining similarity between points in two or more images providing an accurate foundation for motion tracking images [17]. This technique has been used on several works. Cole [6] used this technique to reduce the size of a set of images to which new images were compared. Modegi [31] proposed a structured template matching technique for recognizing small objects in satellite images. There are other methods

of tracking that do not use NCC, including Gradient Descent Search (GDS) and Active Contour Matching [1]. The GDS is based on a first order approximation to image motion and has a restriction that the feature translation is small.

The method of template matching loops the template through all the pixels in the captured image and compares the similarity. While this method is simple and easy to implement, it is the slowest one. [48] This speed problem could be reduced by the application of the metaheuristic population-based optimization algorithms.


1a.  1b.  1c.  1d. 

Fig. 1 Image of the sample part on the assembly conveyor belt, as seen from the overhead camera image. 1a 1b 1c 1d The templates to be detected on the part to predict its orientation for handling by the robot gripper.

In this work, we want to find a reference image in the target landscape image. When the pattern is found in the target image, its rotation angle is determined. To evaluate a candidate solution, the measure of similarity γ between the reference and target landscape image has been proposed. Several similarity measures have been proposed in the literature, such as mutual information and sum of square of differences between pixels [2][6]. In this work, we used the relation in equation 1, considering the degree of similarity between the images.

$$\gamma = \frac{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left[F(x+i,y+j)-\bar{F}_{i,j}\right]\cdot\left[T(i,j)-\bar{T}\right]}{\left\{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left[F(x+i,y+j)-\bar{F}_{i,j}\right]^2\cdot\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left[T(i,j)-\bar{T}\right]^2\right\}^{1/2}} \quad (1)$$

In the equation (1), $F(x,y)$ is the landscape image, $\bar{F}_{i,j}$ is the grey-scale average intensity of the captured image in the region coincident with the template image, $T(x,y)$ represents the template image and $\bar{T}$ is the average intensity of the template image. We have to address that the dimensions of the matrix F is MxN and the size of the template T is mxn. The maximum value of γ is 1, will say that the match between the landscape and the template is perfect. [48].

The FMS layout considered in this work, depicted below, consists of a six axis ABB ERB 6400 robot, a vision camera, and a material handling system- a conveyor belt. The Sony XCG-U100E overhead camera (Figure 2b) is used for identifying the orientation of the part lying on a conveyor belt (Figure 2c).


2a.  2b.  2c.

Fig. 2a. The Assembly Robot 2b. Overhead Camera 2c. Conveyor Belt.

The image captured by the camera is transferred via closed network Ethernet connection to the testing PC. The ABB robot tracks the conveyor belt using a conveyor tracking system which is included in the robot controller. The part orientation information is transferred to the robot gripper via the Ethernet, which then orients itself accordingly to pick the part. The object recognition problem is to find the templates on the parts, such as the one chosen in this case, considering the possible position of the images within the required tact time allocated to the robot assembly cell.


3a.  3b.

Fig. 3a. Gripper orienting to pick the part 3b. Part lifted from conveyor belt

While the simulations provided in this paper are based on real assembly shop data in a company, the actual part details, the assembly cell rates and the gripper construction

details are not revealed due to the proprietary nature of the information.

3. Proposed Methodology

3.1 Bee Colony Algorithm (BCA)

The Bee Colony Algorithm [19] is inspired by the collective behavior of a colony of honeybees working to find food sources around the hive. Although a colony of honeybees has a queen, the control is decentralized rather than hierarchical. The beehive can be understood as a self-organizing system with a multiplicity of agents [24]. A self-organizing system is based on characteristics of positive and negative feedback, random fluctuation as well as the interaction of the system's individuals. The use of preferably good food sources is an emergent property of the beehive.

In BCA algorithm, the position of a food source represents a possible solution to the optimization problem and the nectar amount of a food source corresponds to the quality (fitness) of the associated solution. A colony of honey bees can move itself over long distances and in multiple directions simultaneously to exploit a large number of food sources. The goal of the colony is to achieve good food sources, which depend on some factors such as the distance to the hive, richness or concentration of nectar and easiness of extracting the nectar.

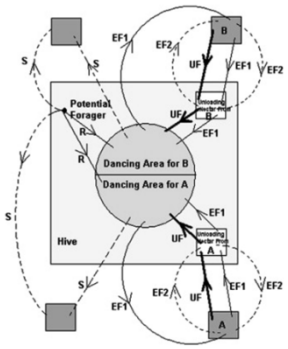


Fig. 4 Behavior of honeybee foraging for nectar (*Adapted from Karaboga et al. 2009*).

A colony of honey bees is classified into three categories; employed bees, onlooker bees and scout bees. All bees that are currently exploiting a food source are known as employed bees. The employed bees exploit the food source and they carry the profitability of the food source back to the hive and share this information with onlooker bees by dancing in the designated dance area inside the hive. Onlooker bees look for a food source to exploit. They watch the dance and choose a food source according to the probability proportional to the quality of that food source. Therefore, good food sources attract more onlooker bees compared to bad ones. Whenever a food source is exploited fully, all the employed bees associated with it abandon the food source, and become scouts. Scout bees will always be searching for new food sources near the hive. The mean number of scouts is about 5–10%. Scout bees can be visualized as performing the job of exploration, whereas employed and onlooker bees can be visualized as performing the job of exploitation.

The main steps of the algorithm are as below: [19]
1: Initialize Population
2: repeat
3: Place the employed bees on their food sources
4: Place the onlooker bees on the food sources depending on their nectar amounts
5: Send the scouts to the search area for discovering new food sources
6: Memorize the best food source found so far
7: Until requirements are met

In BCA algorithm, each cycle of the search consists of three steps: sending the employed bees onto their food sources and evaluating their nectar amounts; after sharing the nectar information of food sources, the selection of food source regions by the onlookers and evaluating the nectar amount of the food sources; determining the scout bees and then sending them randomly onto possible new food sources. At the initialization stage, a set of food sources is randomly selected by the bees and their nectar amounts are determined. At the first step of the cycle, these bees come into the hive and share the nectar information of the sources with the bees waiting on the dance area. A bee waiting on the dance area for making decision to choose a food source is called onlooker and the bee going to the food source visited by herself just before is named as employed bee.

After sharing their information with onlookers, every employed bee goes to the food source area visited by itself at the previous cycle since that food source exists in her memory, and then chooses a new food source by means of visual information in the neighbourhood of the one in her memory and evaluates its nectar amount. At the second step, an onlooker prefers a food source area depending on the nectar information distributed by the employed bees on the dance area. As the nectar amount of a food source increases, the probability of that food source chosen also increases. After arriving at the selected area, she chooses a new food source in the neighbourhood of the one in the memory depending on visual information as in the case of employed bees. The determination of the new food source is carried out by the bees based on the comparison process of food source positions visually. At the third step of the

cycle, when the nectar of a food source is abandoned by the bees, a new food source is randomly determined by a scout bee and replaced with the abandoned one. In our model, at each cycle at most one scout goes outside for searching a new food source and the number of employed and onlooker bees is selected to be equal to each other. These three steps are repeated through a predetermined number of cycles called Maximum Cycle Number MCN or until a termination criterion is satisfied.

An artificial onlooker bee chooses a food source depending on the probability value associated with that food source p_i calculated by Eq. (2):

$$p_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n} \quad (2)$$

where fit_i is the fitness value of the solution i which is proportional to the nectar amount of the food source in the position i and SN is the number of food sources which is equal to the number of employed bees or onlooker bees.

In order to produce a candidate food position from the old one in memory, the BCA uses Eq. (3):

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) \quad (3)$$

where $k \in \{1,2,\ldots., SN\}$ and $j \in \{1,2,\ldots., C\}$ are randomly chosen indexes. Although k is determined randomly, it has to be different from i. ϕ_{ij} is a random number between [-1,1]. It controls the production of neighbour food sources around x_{ij} and represents the comparison of two food positions visually by a bee. As the difference between the parameters x_{ij} and x_{kj} decreases, the perturbation on the position x_{ij} gets decreased, too. Thus, as the search approaches the optimum solution in the search space, the step length is adaptively reduced.

If a parameter value produced by this operation exceeds its predetermined limit, the parameter can be set to an acceptable value. In this work, the value of the parameter exceeding its limit is set to its limit value.

The food source of which the nectar is abandoned by the bees is replaced with a new food source by the scouts. In BCA, this is simulated by producing a position randomly and replacing it with the abandoned one. If a position cannot be improved further through a predetermined number of cycles, then that food source is assumed to be abandoned. Assume that the abandoned source is x_i and $j \in \{1,2,\ldots..,D\}$, then the scout discovers a new food source to be replaced with x_i. This operation can be defined as in Eq. (4)

$$x_i^j = x_{\min}^j + rand[0,1](x_{\max}^j - x_{\min}^j) \quad (4)$$

After each candidate source position $v_{i,j}$ is produced and then evaluated by the artificial bee, its performance is compared with that of its old one. If the new food source has an equal or better nectar than the old source, it is replaced with the old one in the memory. Otherwise, the old one is retained in the memory. In other words, a greedy selection mechanism is employed as the selection operation between the old and the candidate one

3.2 Ant Colony Optimization (ACO)

Ant colony optimization was formalized into a metaheuristic for combinatorial optimization problems by Dorigo and co-workers [27], [28]. One can find ACO metaheuristic application to real-world applications mentioned in the literature such as by Price et al. [29], who have applied ACO to an industrial scheduling problem in an aluminum casting center, and by Bautista and Pereira [18], who successfully applied ACO to solve an assembly line balancing problem with multiple objectives and constraints between tasks.

In ACO algorithms a colony of artificial ants iteratively constructs solutions for the problem under consideration using artificial pheromone trails and heuristic information. Its main characteristic is that, at each iteration, the pheromone values are updated by *all* the m ants that have built a solution in the iteration itself. The pheromone τ_{ij}, associated with the edges i and j, is updated as follows:

$$\tau_{ij} \leftarrow (1-\rho).\tau_{ij} + \sum_{k=1}^{m} \Delta\tau_{ij}^k \quad (5)$$

where ρ is the evaporation rate, m is the number of ants, $\Delta\tau_{ij}^k$ is the quantity of pheromone laid on the edge (i, j) by ant k.

$$\Delta\tau_{ij}^k = \frac{Q}{L_k} \quad (6)$$

if ant k uses edge (i, j) in its tour, and 0 otherwise. In the equation above, Q is a constant, and L_k is the length of the tour constructed by ant k.

In the construction of a solution, ants select the following city to be visited through a stochastic mechanism. When

ant k is in city i and has so far constructed the partial solution s^p ,the probability of going to city j is given by:

$$p_{ij}^k = \frac{\tau_{ij}^\alpha \eta_{ij}^\beta}{\sum_{c_{il} \in N(s^p)} \tau_{il}^\alpha \eta_{il}^\beta} \quad (7)$$

if $c_{ij} \in N(s^p)$, and 0 otherwise. In the equation above $N(s^p)$ is the set of feasible components; that is, edges (i,l) where l is a city not yet visited by ant k. The parameters α and β control the relative importance of the pheromone versus the heuristic information η_{ij}, which is given by:

$$\eta_{ij} = \frac{1}{d_{ij}} \quad (8)$$

where d_{ij} is the distance between the cities i and j.

The pheromone trails are modified by ants during the algorithm execution in order to store information about 'good' solutions. We apply the Ant Colony System (ACS) [9,10], a particular ACO algorithm to the problem on hand, which follows the algorithmic scheme given below:

1: Set parameters, initialize pheromone trails
2: **while** (termination condition not met)
3: *ConstructSolutions*
4: *(ApplyLocalSearch)*
5: *UpdateTrails*
6: **end while**

ACO are solution construction algorithms, which, in contrast to local search algorithms, may not find a locally optimal solution. Many of the best performing ACO algorithms improve their solutions by applying a local search algorithm after the solution construction phase. Our primary goal in this work is to analyze the manufacturing related application capabilities of ACO, hence in this first investigation we do not use local search.

3.3 Particle Swarm Optimization (PSO)

The initial ideas on particle swarms of Kennedy and Eberhart were essentially aimed at producing computational intelligence by exploiting simple analogues of social interaction, rather than purely individual cognitive abilities [34]. The first simulations [20] were influenced by Heppner and Grenander's work [16] and involved analogues of bird flocks searching for corn. These soon developed [9][10] into a powerful optimization method— Particle Swarm Optimization (PSO).

PSO is an optimization algorithm that is based on swarm intelligence principle [9], which are widely used in application domains such as function optimization, neural network training, fuzzy system control and so on at present [33]. It has been proved to be very effective for solving global optimization in various engineering application such as image and video analysis and design and optimization of communication networks. However, most applications in this field are using PSO to train ANN. A direct application of PSO variant in maintenance optimization will be shown in this paper.

3.3.1 Basic PSO Algorithm Description

The Particle Swarm Optimization (PSO) algorithm is a heuristic approach motivated by the observation of social behavior of composed organisms such as birds flocking (Fig.5). A number of simple entities – the particles – are placed in the search space of some problem or function, and each evaluates the objective function at its current location. Each individual in the particle swarm is composed of D dimensional vectors, where D is the dimensionality of the search space.

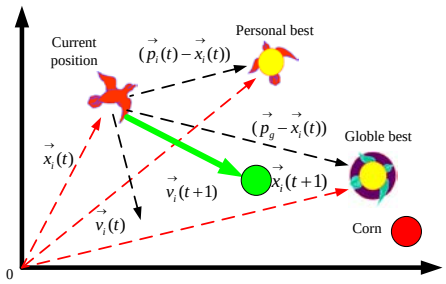


Fig. 5. Bird Flocking of PSO

The current position $\vec{x}_i$ can be considered as a set of coordinates describing a point in space. If the current position is better than any that has been found so far, then the coordinates are stored in the vector $\vec{p}_i$. The value of the best function result so far is stored in a variable that can be called $\vec{p}_g$. The objective, of course, is to keep finding better positions and updating $\vec{p}_i$ and $\vec{p}_g$. New points are chosen by adding $\vec{v}_i$ coordinates to $\vec{v}_i$, and the algorithm operates by adjusting $\vec{v}_i$, which can effectively be seen as a step size. The steps of implementing PSO are shown as follows:

1: Initialize a population array of particles with random positions and velocities on D dimensions in the search space.

2: Loop

3: For each particle, evaluate the desired optimization fitness function in D variables.

4: Compare particle's fitness evaluation with that of its $\vec{p}_i$. If current value is better than that of $\vec{p}_i$, then set $\vec{p}_i$ equal to the current coordinates.

5: Identify the particle in the neighborhood with the best success so far, and assign it to the variable $\vec{p}_g$.

6: Change the velocity and position of the particle according to the following equation:

$$\vec{v}_i(t+1) = \omega \cdot \vec{v}_i(t) + c_1 \cdot r_1\left(\vec{p}_i - \vec{x}_i(t)\right) + c_2 \cdot r_2(\vec{p}_g - \vec{x}_i(t)) \quad (9)$$

$$\vec{x}_i(t+1) = \vec{x}_i(t) + \vec{v}_i(t+1) \quad (10)$$

Where: ω is the inertia weighting; c_1 and c_2 are acceleration coefficients, positive constraint; r_1 and r_2 are the random numbers deferring uniform distribution on [0, 1]; i represents i^{th} iteration.

7: If a criterion is met (usually a sufficiently good fitness or a maximum number of iterations), exit loop.

8: End loop

In PSO, every particle remembers its own previous best value as well as the neighborhood best; therefore it has a more effective memory capability than an algorithm such as the GA. In addition, PSO is easier to implement and there are fewer parameters to adjust compared with GA [8].

3.3.2 Discrete PSO (DPSO) Algorithm Description

The general concepts behind optimization techniques initially developed for problems defined over real-valued vector spaces, such as PSO, can also be applied to discrete valued search spaces where either binary or integer variables have to be arranged into particles [8]. When integer solutions (not necessarily 0 or 1) are needed, the optimal solution can be determined by rounding off the real optimum values to the nearest integer. DPSO has been developed specifically for solving discrete problems. The new velocity and position for each is determined according to the velocity and position update equations given by (8) and (9).

$$\vec{v}_i(t+1) = round(\omega \cdot \vec{v}_i(t) + c_1 \cdot r_1\left(\vec{p}_i - \vec{x}_i(t)\right) + c_2 \cdot r_2(\vec{p}_g - \vec{x}_i(t))) \quad (11)$$

$$\vec{x}_i(t+1) = \vec{x}_i(t) + \vec{v}_i(t+1) \quad (12)$$

In equation (11), the value of velocity is binary or integer because *round ()* function can round off the value.

3.3.3 Improved DPSO (IDPSO) Algorithm Description

DPSO or PSO performs well in the early iterations, but they have problems approaching a near-optimal solution. If a particle's current position accords with the global best and its inertia weight multiply previous velocity is close to zero, the particle will only fall into a specific position. If their previous velocities are very close to zero, all the particles will stop moving around the near-optimal solution, which may lead to premature convergence of algorithm. All the particles have converged to the best position discovered so far which may be not the optimal solution. So, an improved DPSO is proposed here.

In IDPSO, before updating the velocities and positions in every iteration, the particles are ranked according to their fitness values in descending order. Select the first part of particles (suppose mutation rate is α, fist part is (1-α)) and put them into the next iteration directly. Regenerate the rest part of particles (α) randomly. In this project, we can regenerate the positions and velocities according to the following equation:

$$x_{id} = round(rand \cdot (S^{\max}(j) - S^{\min}(j)) + S^{\min}(j)) \quad (13)$$

$$v_{id}(t) = v_{max} - round \cdot (rand \times 2v_{max}) \qquad v_{id}(t) \in [-v_{max}, v_{max}] \quad (14)$$

Because of the characteristics of the flexible manufacturing environment, PSO needs to be discretized. The PSO was modified in order to improve the

optimization effect. Therefore, an improved discrete PSO (IDPSO) was applied in this case.

4. Results and Comparison

4.1 Bee Colony Algorithm

There exist in literature [36] many ways to implement a BCA algorithm. In this paper the bee colony algorithm was implemented in Matlab. For inital tests, we defined the number of employed bees or initial solutions as 100, the maximum number of cycles as 300, and the scout bees as 10% of employed bees. During the search, the stagnation criterion was the non-improvement of the solutions for 10% of the cycles. When stagnation occurred, explosion was performed. All experiments were run to evaluate the object recognition task in digital color and grey images.

The objective of the experiment is to identify the strategies that maximize the average fitness and the number of best solutions that have fitness values greater than 0.95. This value was empirically found and indicates that the object is identified by the algorithm with almost correct coordinates, except by a small tolerance.

The number of food sources in the program can affect to the precision and the velocity of the program. The variation of running time of the program with different number of food sources is shown below to appreciate the differences. In both runs the rule of the same number of employed and onlooker bees have been kept.

4.1.1 With 500 Food Sources

For this experiment, the limit of iterations has been eliminated. This is to avoid be many errors due to the fact that with less food sources there should be more iterations. The Y axis shows the Fitness Values in all plots. With 500 food sources the following was observed:

- The average running time is of 14.70 seconds.
- There has not been any error in the detection of the correct coordinates, all of the templates have reached an NCC higher than 0.52 before 500 iterations.

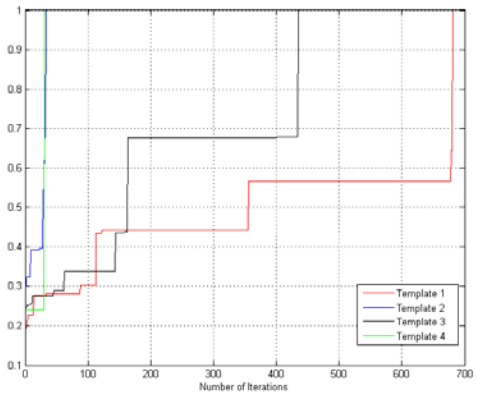


Fig. 6. Testing with 500 food sources

4.1.2 With 100 Food Sources

- The average running time is of 9.19 seconds.
- All of the templates reached NCC value higher than 0.52 before 500 iterations.

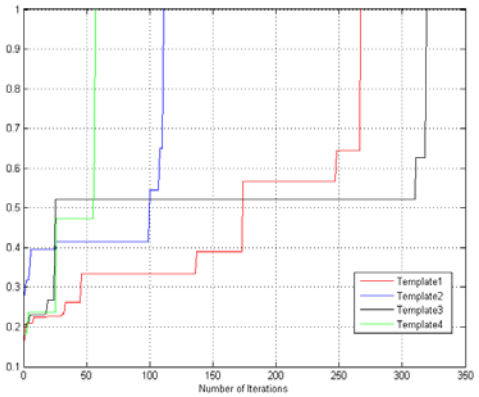


Fig. 7. Testing with 100 food sources

4.1.3 With 10 Food Sources

It was observed that the average number of iterations per template is significantly bigger (4,500) than with more food sources.

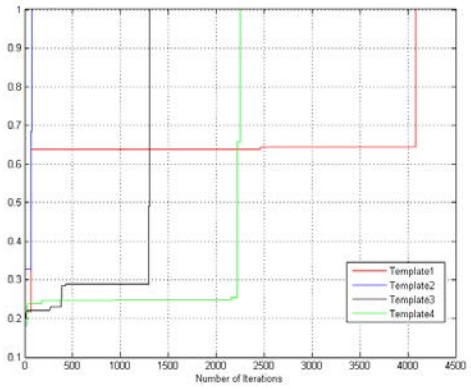


Fig. 8. Testing with 10 food sources

We also noticed that a minimum number of food sources of 10 was the bottom limit required to obtain any resolution with the Bee Colony Algorithm application to the discussed problem.

4.2 Improved Discrete Particle Swarm Optimization (IDPSO)

4.2.1 With 150 particles

To implement IDPSO, a population size of 150 particles was chosen to provide sufficient diversity into the population taking into account the dimensionality and complexity of the problem. This population size ensured that the domain was examined in full but at the expense of an increase in execution time. The other parameters of DPSO and IDPSO were: c1 = c2 = 2.0, ω = 1.2 - 0.8 with linearly decreasing, total iteration = 300 and V $\in$ [-3, 3].

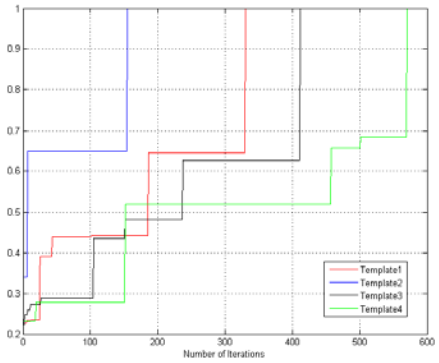


Fig. 9. Testing with 150 particles

The average NCC value of the templates obtained in the experiments was .998 or greater in less than 1000 iterations.

4.2.2 With 500 particles

The average NCC value of the templates obtained in the experiments was .998 or greater in less than 1000 iterations.

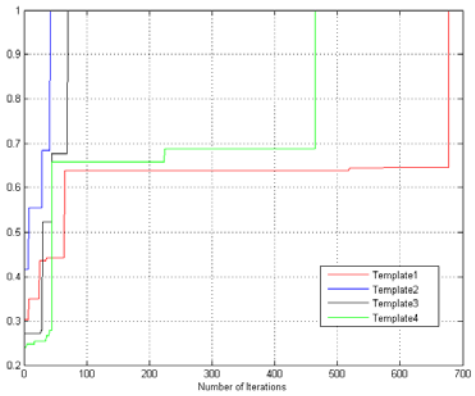


Fig. 10. Testing with 500 particles

4.3 Ant Colony Optimization (ACO)

With ACO we chose the following settings $m = 10, \beta = 2, q_0 = 0.98, \alpha = \rho = 0.1$ and $\Delta\tau_{ij}^{k} = \frac{Q}{L_k}$.[27][28]

The location of the four templates/markers (Figure 11) by the three algorithms is shown in Figure 12. It took an average 8.86 seconds for ACO to find the four templates, and hence the fastest of the three algorithms. Results showed the limits of robustness of the Bee Colony Algorithm, for different food sources. When compared with the results obtained by a particle swarm algorithm [36] for the same problem, they are generally equivalent. The average time taken by ACO, was the closest match to the robot assembly cell takt time of 9 seconds that would be required to establish a lean workflow and reduce machine starvation at the manufacturing facility.

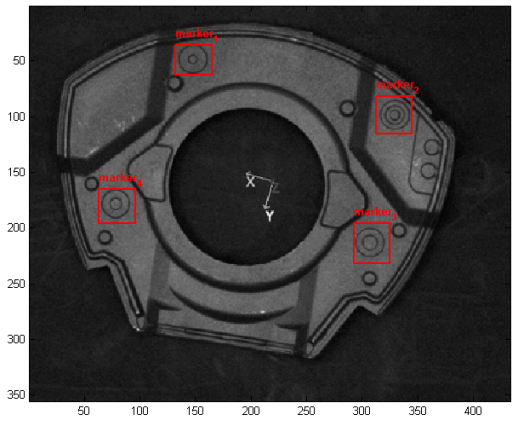


Fig. 11. Templates detected by ACO algorithm

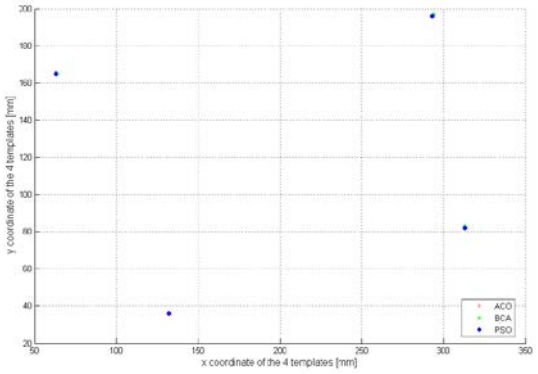


Fig. 12. Coordinates of the 4 templates as solved by ACO, BCA and modified PSO

5. Conclusion

In this paper, the BCA algorithm was tested with variants of Particle Swarm Optimization and Ant Colony Algorithm, and a combination of different strategies, such as generation of scout bees, varying the number of food sources, and explosion of stagnated population. The performance of the Bee Colony Algorithm is good when dealing with images without scaling factor, but this wasn't necessary for our particular manufacturing case study scenario. The choice of algorithms for a manufacturing assembly scenario could vary with the required tact times in the assembly cell, and the production environment such as vibration, dust etc. With real world images, the performance degrades to certain limits, but still finds optimal solution in more than 75% of the cases, and with greater than 10 food sources. It is observed that the computational cost effectiveness of the BCA varies according to the number of food sources chosen. The algorithm can still offer good solutions in the presence of noise within reasonable ranges. Future work will focus on improving the robustness of the algorithm in such situations.

We plan to test other approaches such as comparing the performance of modification such as conventional weight aggregation (CWA) and dynamic weight aggregation (DWA) in multi-objective optimization problems [35], and also compare with other competing evolutionary algorithms, like Genetic Algorithm.

Acknowledgments

Financial support from the AutoCast Consortium and the Norwegian Research Council is gratefully acknowledged.

First Author Rhythm Suren Wadhwa is a PhD student at the department of production and quality engineering, NTNU. She has worked in the Manufacturing Automation industry for five years. Current research interests include assembly automation, optimization techniques, assembly simulation and industrial robotics. She was the president of Society of Women Engineers at the University of Michigan. She has a Masters Degree in Mechanical Engineering and Bachelors degree in Manufacturing Processes Automation Engineering.

Second Author Zhenyou Zhang is a PhD student at Department of Production and Quality Engineering. He is responsible for developing a demo: Intelligent Fault Diagnosis and Prognosis System (IFDPS). Academic interests include Measurement, Mechanical Design, Applied Computational Intelligence (ANNs, ACO, PSO & BCA), Data Ming (Association Rules & Decision Tree), Fault Diagnosis and Prognosis, Condition-based Monitoring and Predictive Maintenance.

Second Author Quan Yu is currently a PhD student at Department of Production and Quality Engineering. He is responsible for studying a 3D Structure Light System. Academic interests include Mechanical Design, Swarm Intelligence (ACO, PSO & BCA), Data Ming and Computer Vision.

Second Author Kesheng Wang is a Professor at the Department of Production and Quality Engineering and head of the Knowledge Discovery Laboratory. Research interests include applied computational intelligence to manufacturing environment.

References

[1] B. D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," presented at International Joint Conference on Artificial Intelligence, Vancouver, 1981.

[2] Brunelli, R.Template Matching Techniques in Computer Vision: Theory and Practice. New York: John Wiley & Sons, 2009.

[3] Chan, T.S., Pak, H.A., Heuristic job allocation in a flexible manufacturing system. Int J Adv Manuf Technol 1(2):69-90, 1986.

[4] Chidambaran, C. and Lopes, H. S., (2010) An Improved Artificial Bee Colony Algorithm for the Object Recognition Problem in Complex Digital Images Using Template Matching, International Journal of Natural Computing Research, Vol1, Issue2, pp.54-70.

[5] Chisman, J.A. Manufacturing cell: analytical set up times and part sequencing. Int J Adv Manuf Technol 1(5):55-60, 1986.

[6] Cole, L., Austin, D., & Cole, L., Visual object recognition using template matching. In Proceedings of the Australasian Conference on Robotics and Automation. 2009

[7] Curkovic, P. and Jerbic, B. (2007) Honey-Bees Optimization Algorithm applied to path planning problem, International Journal of Simulation Modeling, Vol 3, pp. 154-164

[8] Del Valle, Y., Venayagamoorthy G. K., Mohagheghi, S., Hernandez. J., and Harley, R. G., (2008) Particle swarm optimization: basic concepts, variants and applications in power system, IEEE Trans. Evol. Comput., Vol. 2, pp. 171–195.

[9] Eberhart, R. C., and Kennedy, J., (1995) A new optimizer using particles swarm theory. Proceedings of Sixth International Symposium on Micro Machine and Human Science, pp. 39-43.

[10] Eberhart, R. C., Simpson, P. K., and Dobbins, R. W., (1996) Computational intelligence PC tools, Boston: Academic Press.

[11] Evans, H., & Zhang, M. Particle Swarm Optimization for Object Classification. In Proceedings of the 23rd International Conference on Image and Vision Computing (pp. 1-6),2007.

[12] Greenberg, H.H. A branch and bound solution to the general scheduling problem. Int J Oper Res 16:353-361.

[13] Hackel, S. Dippold, P. The bee colony inspired algorithm (BCiA): a two-stage approach for solving the vehicle routing problem with time windows. In Proceedings of the 11th Genetic and Evolutionary Computation Conference (pp. 25-32), 2009

[14] Hambecker, F., Lopes, H.S., & Godoy, W. Jr. Particle Swarm Optimization for Multidimensional Knapsack Problem (pp. 358-365),2007.

[15] Hoitomt, D.J., Luh P.B., Pattipati, K.R., A practical approach to job shop scheduling problems. IEEE Trans Robot Autom 9(1):1-13.

[16] Heppner, H., and Grenander, U., (1990) A stochastic non-linear model for coordinated bird flocks, The ubiquity of chaos, pp. 233–238. Washington: AAAS.

[17] Hii, A. J. H., Hann, C. E., Chase, J. G. and Van Houten, W. E. W., (2006) Fast normalized cross correlation for motion tracking using basis functions. Computer Methods andPrograms in Biomedicine. Vol. 82, No. 2, pp. 144-156.

[18] J. Bautista and J. Pereira, “Ant algorithms for assembly line balancing,” in Proc. ANTS2002, ser. LNCS, M. Dorigo et al., Eds., Springer Verlag, vol. 2463, pp. 65–75, 2002.

[19] Karaboga, D., Akay, B., A comparative study of artificial bee colony algorithm, Applied Mathematics and Computation, 214 (1), 108-132, 2009.

[20] Kennedy, J., and Eberhart, R. C. (1995) Particle swarm optimization. Proceedings of the IEEE international conference on neural networks IV, pp. 1942–1948.

[21] Koren, Y.U.A., Reconfigurable manufacturing systems: Key to future manufacturing. Journal of
Intelligent Manufacturing, 2000. 11(4).

[22] L. M. Gambardella and M. Dorigo. Solving symmetric and asymmetric TSPs by ant colonies. In Proceedings of the 1996 IEEE International Conference on Evolutionary Computation (ICEC’96), pages 622–627. IEEE Press, Piscataway, NJ, 1996.

[23] LaDou, J., (2006) Printed circuit board industry. International Journal of Hygiene and Environmental Health. Vol. 209, No. 3, pp. 211-219.

[24] Lemmens, N., (2006) To bee or not to bee: A comparative study in swarm intelligence. Master’s thesis, Maastricht University, Faculty of Humanities and Sciences. MICCIKAT 06-12.

[25] Lee, S.M., Jung H.J. A multi objective production planning model in flexible manufacturing environment. Int J Prod Res 27 (11): 1981-1992

[26] Lin, Y. H. and Chen, C. H., (2008) Template matching using the parametric template vector with translation, rotation and scale invariance. Pattern Recognition. Vol. 41, No. 7, pp.2413-2421.

[27] M. Dorigo and L. M. Gambardella. Ant Colony System: A cooperative learning approach to the traveling salesman problem. IEEE Transactions on Evolutionary Computation, 1(1):53–66, 1997.

[28] M. Dorigo, V. Maniezzo, and A. Colorni. The Ant System: Optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics – Part B, 26(1):29–41, 1996.

[29] M. Gravel, W.L. Price, and C. Gagn´e, “Scheduling continuous casting of aluminum using a multiple objective ant colony optimization metaheuristic,” European Journal of OperationalResearch, vol. 143, pp. 218–229, 2002.

[30] Mishra, S.K., Performance of Differential Evolution and Particle Swarm Methods on Some Relatively Harder Multi-modal Benchmark Functions. Munich personal RePec (No.1743)

[31] Modegi, T. Small object recognition techniques based on structured template matching for high-resolution satellite images. SICE Annual Conference, 2168-2173

[32] Monkman G., Hesse S., Steinmann R., and Schunk, H. Robot Grippers, (2007).

[33] Pan Hongxia, and Wei Xiuye, (2009) Particle Swarm Optimization Algorithm with Adaptive Velocity and its Application to Fault Diagnosis, 2009 IEEE Congress on Evolutionary Computation, pp. 3075-3079.

[34] Riccardo Poli, James Kennedy, and Tim Blackwell (2007) Particle swarm optimization, Swarm Intelligence, Vol. 1, pp. 33-57.

[35] Parsopoulous, K.E., Vrahatis, M.N. Particle Swarm Optimization Method in Multi objective Problems, SAC 2002, Madrid

[36] Perlin, H.A., Lopes, H.S., & Centeno, T.M. Particle Swarm Optimization for object recognition in computer vision, 2008.

[37] Pham, D.T., Soroka, A.J.,Ghanbarzadeh, A., & Koc, E. Optimizing neural networks for identification of wood defects using bees algorithm. In Proceedings of the International Conference on Industrial Informatics (pp. 1346-1351),2006.

[38] Schulz, J., Mertens P., A comparison between an expert system, a GA and priority for production scheduling. In: Proceedings of the 1st international conference on operations and quantitative management, Jaipur, India, 2.506-513, 1997.

[39] Shankar K,Tzen Y.J. A loading and dispatching problem in a random flexible manufacturing system. Int J Prod Res 23: 579-595. 1985.

[40] Shaw, M.J., Whinston, A.B., An artificial intelligence approach to the scheduling of flexible manufacturing systems. IEEE Trans 21:170-182,1989.

[41] Silva, R.R., Lopes, H.S., & Lima, C.R.E. A compact genetic algorithm with elitism and mutation applied to image recognition (LNCS 5227, pp.1109-1116),2008.

[42] Singh, A. An artificial bee colony algorithm for leaf-constrained spanning tree problem. Applied Soft Computing, 9(2), 625-631. 2009.

[43] Steeke, K.E., Soldberg, J.J., Loading and control policies for a flexible manufacturing system. Int J Prod Res 19(5):481 -490, 1982.

[44] Tereshko, V. & Loengarov, A. collective decision-making in honey bee foraging dynamics. Computing and Information Systems, 9(3), 1-7. 2005.

[45] Toker, A., Kondacki, S., Erkip, N., Job shop scheduling under a non-renewable resource constraint. J Oper Res Soc 45(8): 942-947, 1994.

[46] Wang, K., (2005) Applied Computational Intelligence in Intelligent Manufacturing Systems, Advanced Knowledge International Pty Ltd, Australia.

[47] Wang, K., (2010) Swarm Intelligence in Manufacturing Systems: Principles, Applications and Future Trends, IWAMA (2010)

[48] Wu, C.-H., D.-Z. Wang, et al. (2009). "A particle swarm optimization approach for components placement inspection on printed circuit boards." Journal of Intelligent Manufacturing 20(5): 551-551.

[49] Zhao,X., Lee, M.E., & Kim, S.H. Improved Image Thresholding using Ant Colony Optimization Algorithm. In Proceedings of International Conference on Advanced Language Processing and Web Information Technology (pp. 201-215) 2008.

Electromagnet Gripping in Iron Foundry Automation Part I: Principles and Framework

Rhythm-Wadhwa[1], Terje-Lien[2]

[1,2] Department of Production and Quality Engineering, NTNU
Trondheim, 7051, Norway

Abstract

Robot grippers are employed to position and retain parts in automated assembly operations. This paper presents an overview of electromagnet part handling framework in an iron foundry and an equivalent electromagnet circuit model. The manner in which this whole concept of automated gripping system operates will be discussed in this paper. The material handling system uses machine vision system coupled with conveyor motion and Ethernet communication strategy to assist the material handling system for transporting the foundry parts. The paper provides an overview of the electromagnet principles at play. The electromagnet interaction with the part is the key issue in the robust handling of this automated foundry system. This paper helps in the realization of the concept of automation in an iron foundry, in which the number of published studies is very limited.

Keywords: *Iron Foundry Automation, Electromagnet Gripper Characterestics, Handling Metalcasting*

1. Introduction

Robot grippers are used to position and retain parts in an automated assembly operation. In conventional foundry assembly, such grippers are dedicated to large volume handling of standard parts. The cost of the grippers may be as high as 20% of a robot's cost, depending on the application and part complexity [1]. Electromagnet grippers have several advantages for handling ferrous parts over conventional impactive, ingressive or contiguitive grippers. [2,3] These grippers offer simple compact construction with no moving parts, uncomplicated energy supply, flexibility in holding complex parts and reduced number of set-ups [2]. However, their use is limited to ferrous materials (Iron, Nickel, Cobalt), electromagnet size is directly dependant on required prehension force; residual magnetism in the part when handled when using DC supplies requires the additional of a demagnetizing operation to the manufacturing process. While the choice of material limits application, and demagnetizing is a requirement, the holding force is an important unknown.

Figure 1 illustrates the working principle of an electromagnetic gripper.

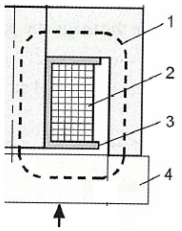


Figure 1: Principle of Electromagnetic Gripper (Adapted from Monkman, 2007)

When placed in contact with an electromagnet, the part provides a flow path for the magnetic flux that completes the magnetic circuit. The force of attraction produced by this circuit holds the part against the electromagnet. During the robot motion, the part tends to slip against the electromagnet surface if the tangential holding force in that direction exceeds the limiting force of static friction for the magnet-part contact. This component of the holding force is in turn dependent on the holding force normal to the electromagnet surface via the coefficient of static friction for the magnet-part material pair [2]. Also, the part flatness and corresponding surface roughness can have an effect on the tangential holding forces. This is important because in practical foundry operations, it is difficult to produce smooth surfaces, which are extremely flat. As a result, the source of variation in the normal holding force observed in the manufacturing plant cannot be easily explained.

Although the users of electromagnets in iron foundries know that factors such as material hardness, surface contact conditions, and electromagnet design influence the holding force, very little scientific work has been reported on this topic. Most of the available literature is of a commercial nature [4,5,6,7].The authors are aware of the gripper mechanisms available in literature [2, 8,9,10,11,12] suitable for handling iron parts, however, there is no study that would adequately clarify the mechanism by which the surface roughness affects the contact forces (normal and tangential) of an electromagnetic gripping head. Recently, Wadhwa et al. reported results from initial survey and experiments on

normal and tangential holding forces by an electromagnet, but the work does not directly address the effect of surface flatness on the holding forces.

Understanding the framework for robust iron foundry prehension will save the need of time consuming costly experimentation to determine the holding forces and/or the adequacy of a given electromagnet in a gripper for assembly handling operation. Moreover, the underlying principles will help the manufacturing engineers understand the theoretical issues to be considered while selecting the electromagnets for part handling in practice and hence avoid damage to the robot and the electromagnet gripper.[8](Figure 8)

This paper presents the system configuration and a theoretical approach to modeling the holding force in magnetic grippers. The approach is termed the magnetic circuit, and utilizes Kirchoff's law for magnetic circuits. A predictive model for the normal holding force combined with the Coulomb friction model for the tangential holding force will enable the selection of optimum operating conditions needed to prevent part slip during part handling and hence avoid damage to the part or the gripper.

2. The Foundry Automation Design Concept

The automation cell was installed between the Fettling and the Assembly area in the Foundry. [14]

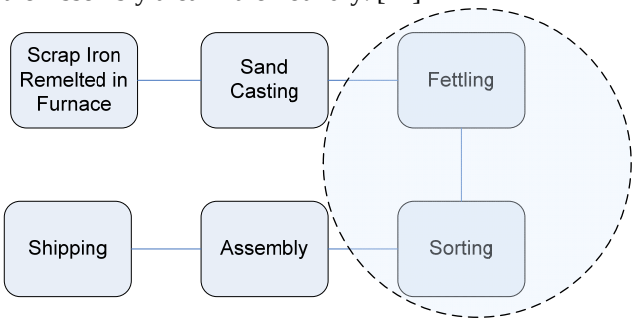


Figure 2: Metalcasting Process

The overall design concept of the Ethernet communication of the material handling system consists of three subsystems, the robot manipulator, material handling system, and the vision system as shown in Figure 5. The intelligent gripper utilized a servo mechanism in which motion controller coupled with the vision system is used for making decisions during the part handling operation.

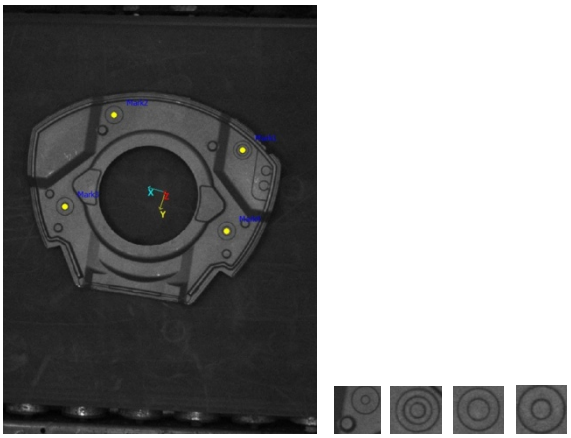


Figure 3: Markers casted on the part

The control system layout shown in Figure 4 shows the control strategy that orients the electromagnet heads according the part orientation. The purpose of the vision system was to recognize the part and extract the orientation. The second purpose of the vision system was to assist in decision making. The part orientation was identified by the markers (Figure 3) which were cast in the part. The vision system conveyed the parts orientation to the robot gripper via the Ethernet and the electromagnets were translated to orient towards the grasping regions.

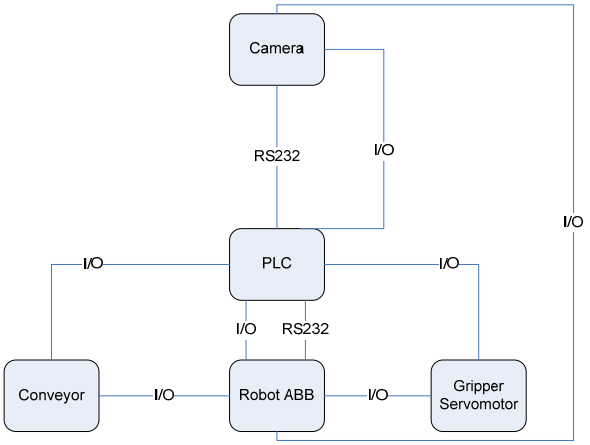


Figure 4: Schematic of Robot Assembly System

3. System Configuration

The six axis ABB ERB 6400 robot used in foundry assembly operation is shown in Figure 5. A Sony XCG-U100E overhead camera was used for identifying the orientation of the part lying on the conveyor belt, which was internally tracked by the robot. The image captured by the camera was processed by Scorpio Vision System (Tordivel AS) and transferred via closed network Ethernet

connection to the Robot. The robot gripper then moved the electromagnets accordingly to pick the part.

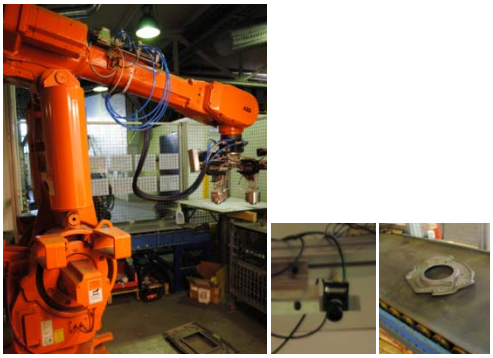


Figure 5: Robot Assembly Cell set up

Figure 6: Gripper picking part from Conveyor Belt.

The part was made of cast iron with the following nominal chemical composition: 3.2 percent Carbon, 2.65 percent Silicon, 0.45 percent Phosphorus, 0.45 percent Manganese, 0.05 percent Sulphur, 0.09 percent Chromium and 0.002 percent Lead. The surface texture of the part taken with IFM 3.1.1 is shown in Figure 7.

While the layouts provided in this paper are based on the real assembly shop data in a company, the actual production volumes and assembly station rates are not revealed due to proprietary nature of the information.

4. Magnetic Circuit Modeling

4.1 Background

The force produced by a magnetic circuit is proportional to the magnetic flux density in the circuit. The amount of flux present depends on the reluctance of the system. The reluctance is low when there is perfect contact between the part and the electromagnet surface. However, part form errors (e.g., surface errors as shown in Figure 7) and roughness lead to an imperfect contact with air gaps between the part and the electromagnet surfaces. Air is a low magnetic conductor and has a permeability ($\mu_0 = 4\pi.10^{-7}\,Hm^{-1}$). This results in large reluctance for elements containing air gaps that in turn can lower the holding force normal to electromagnet surface. Therefore, it is important to model and understand the effects of part flatness and surface finish on the holding forces.

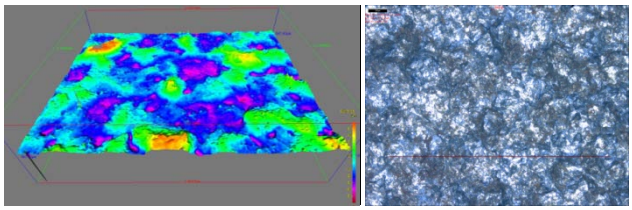


Figure 7: Cast Iron Part Surface (Obtained using IFM 3.1.1.)

4.2 The Magnetic Circuit Approach

The magnetic circuit approach is an analytical method, analogous to electric circuit analysis, for modeling electromagnetic devices [15,16]. Cherry et al. in a classic paper [17], demonstrated the duality between electric and magnetic circuits. The driving force in a magnetic circuit is the magnetomotive force (MMF) $\Im$ which produces a magnetic flux against a coil reluctance $\Re$. The reluctance is defined as:

$$\Re = \frac{l}{\mu A} \qquad (1)$$

Where l is the length of the magnetic flux path, A is the cross section area perpendicular to the flux, and μ is the permeability of the material [15].

For a given MMF and $\Re$, the flux ϕ in the circuit can be found from Kirchoff's law for magnetic circuits. The holding force can be computer using the following simple relation:

$$F = \frac{B^2 A}{2\mu_0} \qquad (2)$$

Where B represents the magnetic flux density in the airgap separating the components, A is the cross section area of the airgap and μ_0 is the permeability of air.

The flux depends on the overall reluctance of the system. The reluctance is low when there is perfect contact between the part and electromagnet. However, part form errors, e.g., roughness (Figure 7), and deviation from flatness give rise to air gaps between the part and gripper. Since actual size and distribution of the airgaps in the gripper-part interface are difficult to determine for a gripper directly in contact with the part surface; it is proposed to model a small uniform air gap that can be

reproduced in an experiment. When the part rests directly on the magnet gripper surface, a uniform air gap equal to the part out-of-flatness error is could be used. It can be assumed here that the reluctance of this air gap is equivalent to the reluctance of the actual contact.

The reluctances proposed in this model include those of the electromagnet, air gaps, part, and the surrounding air medium. The procedure for modeling the reluctances is described next.

4.2.1 Electromagnet Reluctance $\Re_{Electromagnet}$. Figure 1 shows the approximate magnet geometry used to calculate the average cross sectional areas and to simplify the shape path of the flux lines. Note, only half of a cylindrical electromagnet is considered. [2] The magnetic characteristics (B-H curve) could be obtained from the supplier.

4.2.2 Part Reluctance $\Re_{Part}$. The ring-shaped part [ASTM Standards A 773A] has a non-uniform cross section perpendicular to the magnetic flux. The part reluctance is calculated using the following equation [15]:

$$\Re_{Part} = \int \frac{dl}{\mu(l).A(l)} \qquad (3)$$

To evaluate this line integral, a numerical integration scheme can be used. The mean path is such that it is normal to the radial line representing the cross sectional area. The variation in part permeability along the flux path is explicitly accounted for in the calculation of the circuit reluctance.

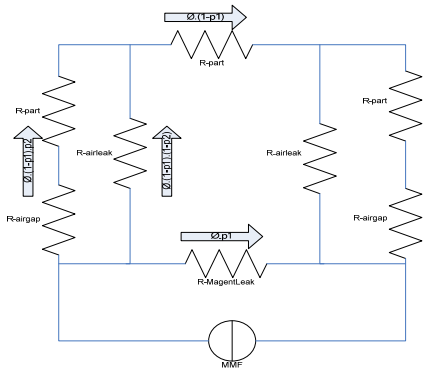


Figure 8: Equivalent Magnetic Circuit of the Magnet-Part System

4.3 Airgap Reluctance $\Re_{Airgap}$. The airgap term applies to the flux lines crossing the magnet-part interface. In reality, the airgap length varies at each point in the interface because of surface roughness and form errors. In this model, an equivalent uniform airgap length is used. The cross-sectional area of the air gap is equal to the magnet-part contact area.

Of the simplifications made above, the use of a mean magnetic flux path is most significant since it implies that the magnetic circuit model cannot predict the distribution of flux in the magnet-part system. However, it can still be used to estimate the total normal holding force and to gain an insight into the effects of magnet and part variables.

4.4 Model Solution. Solution of the magnetic circuit model involves determining the flux flowing through each component of the circuit. This is done using Kirchoff's law for magnetic circuits, which states that the sum of MMF in any closed loop must be equal to zero [15]:

$$\sum_i MMF = \Im - \sum_i R_i\phi_i = \Im - \sum_i H_i l_i = 0 \qquad (4)$$

where index i represents the i th element of the closed loop. H_i is the magnetic field in the i th element, and l_i is the length of the flux path in the i th element. For the magnet used in the gripper, the equation reduces to:

$$\Im + H_{Work}\, l_{Work} + (2B_{Airgap} / \mu_0) l_{Airgap} = 0 \qquad (5)$$

The factor of 2 in the last term accounts for the crossing of airgap twice, once from the N pole to the part and again from the part to the S pole. For the circuit shown in Figure 8, the part holding force is produced by the fraction of flux that crosses the magnet-part air gap, which is given by:

$$\phi_P = \phi.(1 - p_1).p_2 \qquad (6)$$

where p_1 is the fraction of ϕ leaking and p_2 is the fraction of $\phi.(1-p_1)$ entering the airgap and the part. Equation (6) when combined with Equation (2) gives the mechanical force acting on ½ of the model of the part.

5. Conclusions

Foundries, once the low relation of manufacturing engineering, are on their way to automation and undergoing investments in robot installations to increase quality and reduce costs. [7] Material handling is an important portion of foundry plant automation and if not carefully considered, can lead to rapid wear (Figure 9) and high maintenance costs for the robot interface with the part.

The part texture attributes (surface roughness and texture) affect the holding forces of an electromagnet gripper. Future effort in this area will present the effect of these

attributes on normal and tangential holding forces. The results from the magnetic circuit model will be compared with available commercial software and substantiated with experimental analysis.

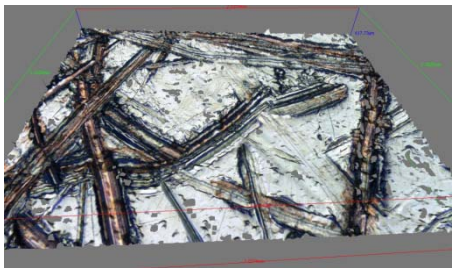


Figure 9: Scratched Electromagnet Surface. (Obtained using IFM 3.1.1.)

References

[1] Pham, D. T. & Tacgin, E. An expert system for selection of robot grippers. *Expert Systems with Applications,1992,*5**,** 289-300
[2] G. J. Monkman, S. H., Steinmann 2007. *Robot Grippers*, Wiley-VCH.
[3] Wadhwa, R.S., Lien,T., Monkman, G.J. Robust Prehension for ferrous metalcasted product families, Proceedings of MITIP 2011
[4] *Anonymous*, New shelf robot saves vital space in the foundry environment, The Industrial Robot. Bedford: 2006. Vol. 33, Iss. 2; p. 145
[5] *Anonymous*, The Castings Center selects STRIM, Euclid, and Prelude Software, The Industrial Robot. Bedford: 1996. Vol. 23, Iss. 6; p. 6
[6] Wetzel, S. GM's Iron Finishing Automation, Modern Casting, 2008; 98,1 ABI/INFORM Complete pg.38
[7] Rooks, B.W. Robots at the core of foundry automation, The Industrial Robot, 1996; 23,6 ABI/INFORM Global pg.15
[8] Hesse,S; Schmidt, H; Schmidt, U.: *Manipulatorpraxis*, Vieweg Verlag, Braunschweig/Weisbaden 2001
[9] CHEN, F. Y. 1982. Gripping mechanisms for industrial robots: An overview. *Mechanism and Machine Theory,* 17**,** 299-311.
[10] Luo, R. C. Year. Automatic Quick-Change Gripper Finger For Assembly Automation. *In*, 1984.215-224.
[11] Salisbury, J. K. & Craig, J. J. 1982. Articulated Hands - Force Control And Kinematic Issues. *International Journal of Robotics Research,* 1**,** 4-17.
[12] Schmidt, I. 1980. Flexible moulding jaws for grippers. *The Industrial Robot.*
[13] Perry, M.P.: Low Frequency Electromagnetic Deisgn, Marcel Dekker, 1985
[14] Campbell, J, *Castings*,2003,Butterworth-Heinemann
[15] Hoole, S.R, *Computer Aided Analysis and Design of Electromagnetic Devices*, 1989
[16] Law, J.D., Modeling of Field Regulated Reluctance Machines, PhD Thesis, University of Wisconsin-Madison,1991
[17] Cherry, E.C.: The duality between interlinked electric and magnetic circuits and the formation of transformer equivalent circuits, Proc. Phys. Soc., 1949, 62, p.101

Acknowledgments

Financial support from the AutoCast Consortium and the Norwegian Research Council is gratefully acknowledged.

First Author Rhythm Suren Wadhwa is a PhD student at the department of production and quality engineering, NTNU. She has worked in the Manufacturing Automation industry for five years. Current research interests include assembly automation, optimization techniques, assembly simulation and industrial robotics. She was the president of Society of Women Engineers at the University of Michigan. She has a Masters Degree in Mechanical Engineering from University of Michigan, and Bachelors degree in Manufacturing Processes Automation Engineering.

Second Author Terje Kristoffer Lien is a Professor in Manufacturing Automation and Robotics at the department of Production and Quality Engineering, NTNU. He has been active in the deveopment of cellular manufacturing systems. His work has attracted international interest, in particular the use of force feedback as a programming tool, and as an enhancment of the control of robots used for grinding operations.

Conception and Use of Ontologies for Indexing and Searching by Semantic Contents of Video Courses

Merzougui Ghalia[1], Djoudi Mahieddine[2] and Behaz Amel[3]

[1] Departement of computer science, Faculty of Science, Batna University, (05000) Algeria

[2] Laboratory XLIM-SIC and IRMA a Research Group, UFR Sciences SP2MI, University of Poitiers Teleport 2, Boulevard Marie et Pierre Curie BP 30179 86962 Futuroscope, Chasseneuil Cedex- France

[3] Departement of math, Faculty of Science, Batna University, (05000) Algeria

Abstract

Nowadays, the video documents like educational courses available on the web increases significantly. However, the information retrieval systems today can not return to the users (students or teachers) of parts of those videos that meet their exact needs expressed by a query consisting of semantic information. In this paper, we present a model of pedagogical knowledge of current videos. This knowledge is used throughout the process of indexing and semantic search segments instructional videos. Our experimental results show that the proposed approach is promising.

Keywords: *video course, ontology, OWL, conceptual indexing, semantic search, vector method adapted.*

1. Introduction

The e-Learning is largely based on multimedia materials and particularly on videos. Many institutes, schools and associations on the web diffuse video lectures on scientific conferences, seminars and thesis dissertations or habilitations (e.g. INRIA, ENS, Aristotle…). Some Universities (or virtual campus) diffuse on the Internet their lectures as audio or video (are cited as an example: MIT, Berkeley, Strasbourg, MedNet and Lausanne). In addition, university lectures are grouped in thematic portals such as WebTV Lyon3 or SciVee (one of many examples of sites dedicated to science videos). These videos are recorded in different formats: i.e. video streamed (or podcast) or structured multimedia documents (where video and presenter's voice are synchronized with slides), and this for a live broadcast or delayed.

While these video documents are more accessible to their richness and semantic expressiveness and their numbers are growing more and more, their treatment remains problematic. In particular, the search for relevant video sequences according to criteria related to the semantic content is not trivial. This can affect the learner while revising it or the researcher (or teacher) who wants to reuse a portion of a video for him. It is often more convenient for a user (learner or teacher) to use semantic information in its query (scientific concepts) to get the most relevant answers. Therefore a process of indexing and searching by the semantics of this type of video should be set up.

Before reaching this stage, it should be noted that it is virtually impossible to achieve the semantic level, starting from a low-level analysis of video content. Interpretations of the contents of a video, which are semantically richer, make the task of the indexer more complicated than the case of a by keyword indexing. This is because he must choose the best index to describe content very rich in information. One meets the same difficulty in the research process. So we must first develop models capable of describing and modeling the semantic content of these videos in order to facilitate access, reuse and navigation by semantics.

In this context, the processing of video content using techniques of knowledgebase is an interesting idea. In the perspective of Semantic Web, which is becoming a basis for distance learning environments, the ontology provides a rich semantic better than any other method of knowledge representation [1]. In a teaching platform, the precision of a search for educational content can be improved if based on the conceptual vocabulary defined in ontology while avoiding the ambiguities in terminology and allowing inferences that reduce noise and increase relevance.

Our work aims to develop ontological models to form a conceptual vocabulary shared between the teachers and learners. We will use this vocabulary in the

annotation of videos from university lectures. Then, we seek to develop a system for indexing and searching the semantic content of video segments, based on their ontological annotation to overcome the lack of such a tool actually.

In Section 2, we present different approaches to video indexing, namely, classical / semantic, automatic / semi-automatic, low level / high level. Section 3 will come later, and will present some work using ontologies for indexing documents in the two areas of interest, namely the e-learning and audiovisual. Then we describe in Section 4 our approach which is divided into five stages: ontological modeling of semantic content of video courses, their annotation based on models developed, conceptual indexing, conceptual research and finally experimentation. We conclude by specifying the limits and prospects of our approach.

2. Semantic Indexation of Video

The indexing of the video document is difficult and complicated. This type of material does not decompose to easily identifiable units as is the case for text document. It is therefore necessary to have tools able to segment, to describe and to annotate the content [2]; this is the task of annotating video document.

Charhad has defined video annotation as the process by which text informations (or other) are associated with specific segments of video material to enrich the content. This information does not modify the document but is just mapped to it. The annotation is often considered to be a laborious task that requires human intervention; however, it remains in high demand for describing the semantic content of a video.

Several standards are used to describe or to annotate multimedia content, such as Dublin Core [3] and MPEG7 [4], using a defined list of attributes such as creation date, authors, image resolution, etc... Dublin Core is used to describe the data cataloging bibliographic records. MPEG 7 is used to describe, in a fine low-level, visual and sound elements of an audiovisual document (such as texture, dominant color ...). But Troncy in [5] found that the descriptors of the latter are too low to accommodate all the needs of semantic description. These standards are far from satisfactory because each provides few mechanisms of knowledge. Therefore ontologies can supplement them in order to access the level of multimedia semantics.

Indexing based on ontologies to represent the documentation granules is called semantic indexing. It consists in choosing the set of concepts and instances of ontology as a representation language of the documents. The granules are then indexed by concepts that reflect their meaning rather than words quite often ambiguous. One should use an ontology reflecting domain or domains of knowledge discussed in the document collection [6].

Hernandez identified two steps to semantic indexing [6]. The first step is to identify the concepts or instances of the ontology in the granules also called conceptual annotation. The second step is to weight the concepts for each document based on the conceptual structure from which it originates.

In what follows, we will cite some works that use ontologies for indexing video documents corresponding to two areas: e-learning and broadcasting.

3. Existing Works

The use of ontology in the context of indexing has grown in recent years in various fields; we cite two that interest us: the audiovisual sector [2], [7] and [8] and the field of e-Learning [9], [10] and [11]. In this latter, several studies are based on the general idea of indexing document fragments on the basis of different kinds of ontologies: i.e. ontology document structure, domain ontology or pedagogical ontology (figure, formula, equation...), to reuse them to compose more or less automatically new resources.

For example, the IMAT project [9] is turned to the use of ontology for indexing. The handling of documents requires the use of an ontology called 'document' which adds to the traditional domain and pedagogical ontologies. The ontology of the course is divided into three sections describing the content, the context and the structure. The content is the domain ontology and the other two parts are related to the pedagogical aspects (structuring the chapters, nature of the parties, etc.).

The project MEMORAe [10] describes organizational memory training based on two ontologies. The first is domain ontology which describes the concepts of the training: individual (student, teacher...), documents (book, web page ...), educational activities (courses, TP...). The second is an application ontology that specifies all the concepts useful for specific training such as algorithms or statistics.

The project Trial Solution [11] consists in taking each educational resource and breaks it down into learning objects 'OP'. Each OP is represented by its semantic content and its relationship with the other OP and metadata that concern. An annotation tool was developed by the project that indexes each node by metadata and by the terms of a thesaurus. May be mentioned other similar works in the field of e-learning [12], [13] and [14].

In the audiovisual field, we cite the work of [7] which articulates a specific conceptual knowledge of a domain through ontology in the context of indexing of audiovisual materials on the theme 'Sport TV emission' In this work, there are two ontologies:

- An audio-visual ontology to standardize the meaning of terms commonly used to describe the structure and format of audiovisual materials. For example a schema that indicates that a sports magazine is like sports program and that always begins with start sequence plateau, followed by a number of sequences that are either plateau sequence or a sequence launch-plateau-report and ends with a sequence end.
- The second is the domain ontology that models the concepts of a particular sport which is his example cycling (tour de France, sports magazine, etc.)...

The work of Isaac [8] is based on the semantic description of the content of television programs on the theme of medicine. It combines multiple ontologies namely: ontology of AV and thematic ontologies related medical fields (MENELA which describes the field of coronary disease and includes concepts related to cardiac surgery correspondent to the topic of the corpus, GALEN contains concepts related to all medical fields).

In [2], Charhad proposed a model for the representation of semantic content of videos. This model allows synthetic and integrated consideration of information components (image, text, sound). He developed a number of tools to extract concepts (name of a person, a geographical place or an organization) and one for detection and recognition of the identity of the speaker which is based on the analysis of automatic transcription of speech in a video.

We also cite some recent studies, [15] and [16], on the treatment of the semantic content of video representation of the field of e-learning. Dong et al. [15], offer a model of multi-ontology annotation of multimedia documents. But they focus in their paper on video presentations of lectures, seminars and corporate training. Each segment is annotated from a Multimedia Ontology (OM) and several domain ontologies. The 'OM' ontology is based on the standard MPEG7, but focuses on the aspect of content description. It contains three types of classes or concepts: multimedia concepts (image, video, audio, video segment, etc…), non-multimedia concepts (agent, place, time, etc…) and descriptor concepts of domain ontologies (such as Gene Ontology 'GO '). We note that the pedagogical aspect is missing in the annotation in this work.

First, we must state that in e-learning, the course material is available on the web in two categories of documents: static multimedia documents (such as web page, pdf, doc, etc…) and dynamic or temporal documents (such as video, audio or SMIL [17]). The number of documents of the second category continues to grow while indexing jobs in the field of e-learning mentioned above are based only on documents of the first type.

Second, the temporal nature of such documents creates a number of constraints on their management. Indeed, the specificity of these documents is to be temporal objects. Inherently this temporality does not present itself and can not be stored and this has several consequences. One of them is the imposition of rate of reading the document, if the video is an hour, it takes an hour to see it and if the information sought begins at the 12th minute and lasts 10 minutes, then you have to wait all this time or scroll through the first 11 minutes to find her.

Third, the video courses posted on the web are annotated in general metadata (format, creation date, author, title, keywords, and sometimes abstract). Note that the use of free text (keywords, abstract), to describe the content, prevents the control of the description's semantic and this severely limits the possibilities of reasoning.

Search engines currently available do not allow a search by the semantic content of a sequence (or segment) in a video course because it is not logically structured, nor semantically described. There are no tools that offer this possibility so far.

So our contribution is in the context of offering the community of e-learning (learner or teacher) a system that helps to search the semantic content of instructional video segments.

4. Approch

In a context of video courses information seeking by semantic content, modeling is an important and necessary task from which the index will be formulated and the by which research process will be more efficient and more accurate. Our approach comes as a modeling and indexing of pedagogical video courses and research through the semantics of the segments in such videos.

At the theoretical level, our contribution consists in the proposal and the construction of two types of ontologies, one for the pedagogical structuring of a video course and the other for describing the semantic content of its various granules. Both ontologies will be used in the phase of conceptual annotation.

At the experimental level, our contribution consists, first, in the conceptual annotation of a corpus of video course about continuous professional training broadcast on the Web from the University NETTUNO under the project

MedNet'U. Through this project MedNet'U (Mediterranean Network of Universities), satellite channels Rai Nettuno Sat forwarded academic lessons on professional training arguments in four languages: Italian, French, English and Arabic.

The annotation is done on ten video lessons from the course 'data structure and algorithm' (in French). Then, our contribution consists in the implementation of the prototype IRSeCoV: a system of indexing and semantic searching of pedagogical video
segments through conceptual annotations associated with the corpus and we follow by an experiment.

4.1 Construction of Ontologies

We need two ontologies to model the content of courses in video format. The first will be built for pedagogical structuring of a video course and will be called pedagogical ontology of the video course. The second is the ontology of the domain of teaching and, as its name suggests, it will model the knowledge of a subject area (a teaching modulus) for a deeper semantic description of this type of course. We begin by describing the latter.

4.1.1 Ontology of the Domain of Teaching

A domain or area of teaching is a single module within training. A module addresses or teaches one or more concepts. A concept can be broken down into several concepts; it may depend on one or more concepts as may be the prerequisite of one or more concepts as well. So, three types of relationships may exist between two concepts: '*is_decomposed_into*', '*depends*', and '*is_prerequisite*'. Note that '*is_prerequiste*' has the characteristic of transitivity while '*depends*' is symmetrical and '*is_decomposed_into*' is anti-symmetric.

It is noted that the exploitation of the characteristics of these relationships can generate or infer instances not found in the basis of the original facts.

Consider the example of teaching domain "*data structure*" which discusses the concepts: function, parameter, parameter passing by value, list, pointer and recording. Instances of relationships that can exist between these concepts are represented as follows:

- *is_decomposed* (function, parameter). The function is composed of parameter.
- *depends* (function, parameter_passing_by_value). Sinc e this relationship is symmetrical, the inference system can deduce the next instance:
- *depends* (parameter_passing_by_value, function).
- *is_prerequiste* (pointer, list) and
- *is_prerequiste* (list, tree) => *is_prerequiste* (pointer, tree). The relation is transitive.

Fig. 1 shows this ontology in the form class diagram.

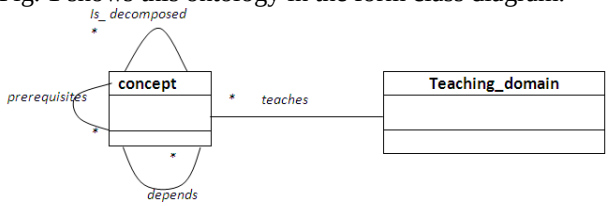


Fig. 1. Domain Ontology of Teaching.

We can mention the existence of instances of the class concepts that are identical. We cite as an example: loop and repetition's instruction, address and pointer, record and structure, two-dimensional table and matrix, parameter and attribute, etc... This semantics is specified in OWL [18] by the property 'sameAs' between individuals.

An instance of this ontology creates domain ontology to teach D (a specific module). It can be manual or semi-automatically. In the second case, this process makes use of language engineering tools (such as LEXTER) for the extraction of candidate terms from one or more textual course materials in a particular field D. These terms represent the extension of the class concept and should be selected, sorted by an expert in the field and organized hierarchically according to the relation '*is_decomposed_into*'. Then, the semantics of the domain is refined by the precision of identical instances of the class concept and the relations of the two bodies '*is_prerequisites*' and '*depends*' that can exist between different concepts.

For our part, we manually created an ontology for the module '*Data Structure*' with the publisher '***protégé 2000***'. Below one can find an excerpt from the OWL code generated by this tool. We used the French language because the video lessons we annotate are in French.

```
<?xml version="1.0"?>
<rdf:RDF
    xmlns=http://www.owl-
ontologies.com/Ontology1277939276.owl#
        ……
<Teaching_domaine rdf:ID="structure_de_donnee">
<teachs>
 <concept rdf:ID="instruction">
  <is_decomposed rdf:resource="#affectation"/>
  < is_decomposed
   rdf:resource="#instruction_de_controle"/>
 <is_decomposed
  rdf:ID="#instruction_de_repetition"/>
 <concept rdf:ID="boucle"/>
   <owl:sameAs rdf:resource=
    "#instruction_de_repetition"/>
 </concept>
 <concept rdf:ID="passage_parametre_par_valeur">
   <depends rdf:resource="#fonction"/>
 </concept>
 <concept rdf:ID="pointeur">
   <prerequisites rdf:resource="#liste"/>
 </concept>
```

```
…
</teachs>
…
```

4.1.2 Pedagogical Ontology of a Video course

A video course is presented in one or more video lessons. Each video lesson (it can be a chapter or sub chapter) is divided or segmented into several temporal segments. The segment corresponds to the explanation of one or more slides with the same title. So the segment in this case must represent an idea or a subject or unit of interest which will be returned by our search system.
A segment or a slide contains one or more pedagogical objects 'POb' (or Learning Objects). It can be a definition, an example, an exercise, a solution_exercice, an illustration, a rule, a theorem, a demonstration, etc....
While viewing some video lessons from the corpus that we chose, we found that a slide contains a definition of a concept followed by a small example. We also noticed that one example can be presented in two or three slides with the same title; we opted for this manner of structuring segments.
This POb concerns one or more concepts of a teaching domain. The relation '*concerns*' manages the alignment of two ontologies: i.e. pedagogical ontology of the video course (POV) and domain ontology of teaching (DOT). To do this, there is a need to import the second (DOT) into the first (POV) (see Fig. 2).
A question: why a POb concerns a number of concepts and not one. If we take the example of the video lesson with the title 'functions', where there is a slide with an exercise on the use of a table as a parameter of a function, we see that the POb-type 'exercise' concerns the three concepts (underlined) of the domain ontology of teaching 'Data Structure' (to our knowledge, there is no search engine that responds to a request of this type).

4.2 Annotation Process

Some authors, such as Charhad, consider or call the annotation phase as assisted or manual indexing. This phase describes the pedagogical video documents by considering two aspects: one is pedagogical and identifies the components of the educational structure of the document (slide, learning object type definition, example ...) and the second is thematic and describes each element as a concept in the field.
Troncy and Isaac used the tool SegmentTool, while Charhad used VidéoAnnex. In our case, we have developed a new tool for segmentation and annotation of video course called OntoCoV and based on ontologies we created.

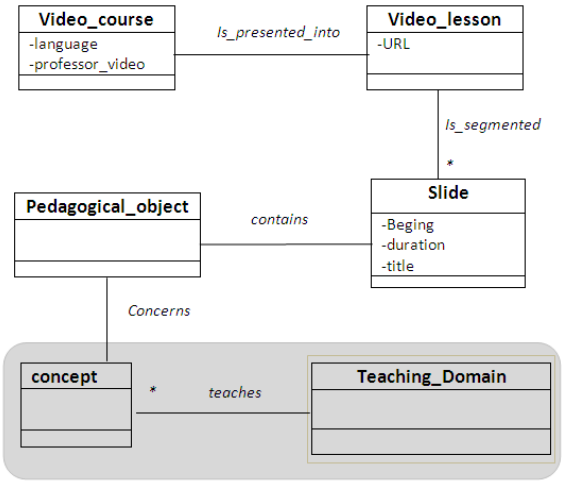


Fig. 2. The Pedagogical Ontology of the video course.

The description or annotation via OntoCoV begins with the localization or temporal segmentation into entities regarding an indivisible concept. It comes to identify segments in time, where each corresponds to exposure of one or more slides with the same title. Then, each segment must characterize its pedagogical structure. These two steps generate the instantiation of the ontology of video course. Next, a description of the semantic content of each segment is made by the association of concepts of the ontology of a teaching field that is particular to different pedagogical objects (POb).
This ontology must have a close relationship with the video lesson being annotated. So OntoCoV gives its user the possibility to integrate ontology of a particular area. This ontology is presented in our tool as a tree graph, which allows the user to quickly browse and select, at all levels (hierarchy of concepts), the concept that seems pertinent for its indexing. At the end, the system will generate all the annotations in the operational language OWL. These annotations provide a basis of facts that will be exploited in subsequent phases.

Fig. 3. The interface of OntoCoV tool.

Fig. 3 shows the interface of the tool OntoCoV which is divided into two regions:

(a) Area to watch the video with buttons to play the video, stop it, create a segment, etc...

(b) Region to display all the segments slide built.

A small separate window appears by clicking on a segment of the area (b). It contains a list of pedagogical objects forming the selected segment. We can describe each POb by associating a list of concepts of the ontology of teaching field already built into the tool.

After the annotation of a video lesson of the course 'data structure and programming techniques', the tool will generate the following OWL code:

```
…
<video_course rdf:ID="structure_de_donnee">
 <is_presented_into>
  <lesson_video rdf:ID="fonction">
        <URL
rdf:datatype="http://www.w3...#string">
       http://.../fonction.wmv </URL>
   <is_segmented rdf:resource="#slide_2"/>
   <is_segmented rdf:resource="#slide_3"/>
   <is_segmented rdf:resource="#slide_7"/>
  </lesson_video>
 </is_presented_into>
<langage rdf:datatype="&xsd#time">frensh</langage>
…
</cours_video>
<slide rdf:ID="slide_2">
 <Duration rdf:datatype="&xsd#time">00:03:22
  </Duration>
 <Beging rdf:datatype="&xsd#time">00:02:01
  </Beging>
 <Title rdf:datatype="&xsd#string">introduction au
  function</Title>
 <contains>
  <POb rdf:ID="definition_1">
     <concerne rdf:resource="&p1#adresse"/>
     <concerne rdf:resource="&p1#fonction"/>
  </POb>
 </contains>
 <contains>
  <POb rdf:ID="exemple_1">
   <concerne rdf:resource="&p1#valeur_retournee"/>
   <rdfs:comment rdf:datatype="&xsd#string">
    differents type de valeurs retournée
   </rdfs:comment>
  </POb >
 </contains>
</slide>
…
```

4.3 Conceptual Indexing

Once the concepts of both ontologies have been identified in the temporal segments, we move to the phase concept's weighting for each slide.

In this part of our work, we present an index structure that can pose queries on temporal segments of video document. For this we use the vector model of Salton [19] while adjusting the calculation of the weight TF_IDF (Term Frequency_Inverse document Frequency) for our needs, drawing on the works of [20] and [21].

It is proposed that the document for a video lesson is no longer represented by a vector but a matrix of concepts and temporal segments. Since the segments are described by concepts instead of words, we compute the weight of concepts with respect to segments in which they appear. Thus we define the new formula CF_ISDF (Concept Frequency_Inverse Segment and Document Frequency), as follows:

$$CF-ISDF(c,s,d) = CF(c,s,d) \times \\ \times ISF(c,s,d) \times IDF(c,d)$$

$$ISF(c,s,d) = \log\left(\frac{S_d}{SegF(c,s)}\right)$$

$$IDF(c,d) = \log\left(\frac{S_d}{SegF(c,s)}\right) \quad (1)$$

$CF(c,s,d)$: The number of occurrences of concept c in the segment s of the document d.

D : Set of all documents (video lessons) of the corpus.

S_d : Number of segments in the document d.

$SegF(c,s)$: Number of segments in the document d in which the concept c appears.

$DF(c)$: Number of documents containing the concept.

This formula allows us to balance the concept not only by its frequency in the segment s on a document d, but also its distribution in the document ($ISF(c, s, d)$); this last measure represents the discriminatory strength of a concept c in the document d. The distribution of the concept in the corpus ($IDF(c, d)$) is also important. If a concept appears in several documents, it is less representative for a given document with respect to another concept that appears only in only one document. It is the discriminatory strength in the corpus.

4.4 Conceptual Search

Now we come to the search phase which is also called the interrogation phase. It includes:

- formulating the need for information through query,
- translating the query into an internal representation defined by a query template,
- comparing the request to document's indexes in the corpus by the correspondence function,
- presenting the results in order of relevance.

Formulation of the query: The corpus of video courses covers several subjects or areas of teaching. For each TD we associate an ontology according to the model developed above. These ontologies, which were used during the annotation, will be used to help the user formulating its query. Our system provides an interface for visualization and exploration of an ontology of a particular TD, chosen by the user, to guide him to browse the tree of this ontology and giving him the opportunity to choose the concepts of his query (see Fig. 5).

The reason that led us to choose this way of query formulation is twofold:

- the user has indeed difficulty to specify his need and to express it,
- one must remove ambiguities and improve the precision and recall of our system.

Query template: For each query (as for segments), we associate a vector. We can assign a weight to the concepts of the query. We assign the value 1 when the concept is present and 0 otherwise.

Correspondence Function: We adapt a measure of pertinence from classical vector model. The relevance of a query *Q* over a segment *S* of the document *D* is:

$$pertinence(S_D, Q) = \cos(V_{S,D}, V_Q) \quad (2)$$

$V_{S,D}$ and V_Q are respectively the vectors of weight of concepts in segment *S* of document *D* and query *Q*.

The Search Results: is a list of references to the most pertinent segments, viewed in order of pertinence in a page coded in HTML + time. Hence the user can see and read a selected video segment on the same page.

4.5 Prototype and Experiment

To evaluate our approach, we implemented a prototype called IRSeCoV (Abbreviation of French translation of: Indexing and Semantic Search in Video Course). Our system aims to allow a more accurate and relevant search of pedagogical video segment. It has several components (see Fig. 4 and Fig. 5) which allow it to be modular.

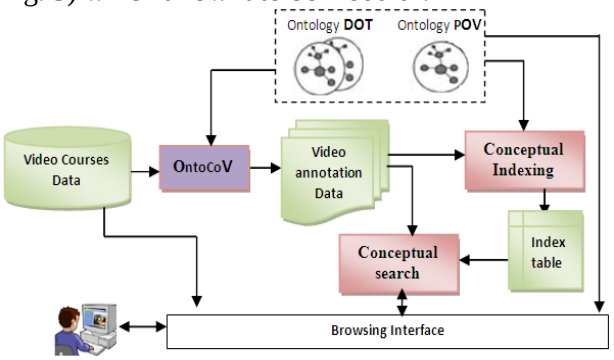


Fig. 4. The General architecture of IRSeCoV system.

To evaluate our system, we have two approaches:

- The first concerns the evaluation of the structure of the index. It comes to calculate the time of indexing, the storage space of the index relative to the size of the corpus, the time of constructing the ontology and the time of corpus annotation. Calculating of time of constructing the index does not assess the value of the index.
- The second concerns the evaluation of the relevance of the index by testing its impact on research using traditional measures of relevance (recall and precision).

Initially we tested the weight of some concepts taken from the index table that was generated by our system.

The experimentation of our system was done on a corpus (annotated by OntoCoV) of 9 video lessons (from 25) of the module 'data structure and programming techniques' which was given during a continuing professional formation by the virtual university NETTUNO under the project MedNet'U. The following table shows some concepts and the list of segments in which they appear. The segment is defined by the document number (video lesson) and the segment number (slide) in this document.

Table 1: List of concepts associated with segments

Concepts	List of segments
Pointeur	{(D_1,S_7), (D_4,S_1), (D_5,S_5), (D_5,S_{12}), (D_6,S_2), (D_8,S_9), (D_8,S_{14}), (D_9,S_4)}
Parametre_formel	{(D_4,S_2)}

CF-ISDF(Pointeur,S_2,D_6) = 1×log(13/2) ×log(9/6)
= 0,7589.

CF-ISDF(Pointeur,S_9,D_8) = 6×log(16/11) ×log(9/6)
= 0,9110.

CF-ISDF(Pointeur,S_{14},D_8)= 1×log(16/11) ×log(9/6)
= 0,3038.

CF-ISDF(Parametre_f,S_2,D_4) = 1×log(14/2) ×log(9/1)
= 4,2756.

- If a concept appears in two segments of the same lesson, the frequency determines the most pertinent segment. See the concept 'pointeur' in the two segments (D_8,S_9) and (D_8,S_{14}).
- If a concept appears in segments from different lessons, then the discriminatory strength ISF determines the most pertinent segment. See the concept 'pointeur' in the segments (D_6,S_2) and (D_8,S_{14}).

If a concept appears in one or two segments at most in the corpus, it will have a great weight because of the discriminatory value IDF; as can be seen with the concept ‘parametre_formel’.

The user can specify in his query the pedagogical object with the concepts he seeks. The system returns a list of segments sorted by pertinence. For each segment, it displays the name of the lesson, its beginning, its duration, its title and more importantly, the pedagogical objects included in the segment with a comment. The user can thus select the segments as needed (see Fig. 5).

Fig. 5. System interface IRSeCoV.

We plan to expand the testing of our system on a corpus of video courses from different teaching field to assess its relevance by calculating recall and precision.

5. Conclusions

We have presented in this paper an approach of searching by the semantic content of pedagogical video segments using ontologies. We built two ontologies, the first structure, pedagogically, a video course and the second models the knowledge of a teaching field.
We realized a new tool called OntoCoV that generates the annotation of video lesson in OWL-based on ontologies.
Then we detail the indexing and the conceptual searching of all annotated video course by adapting the vector method. We have defined a new formula CF-ISDF to calculate the weight of a concept in a video segment. To implement this approach, we developed the prototype IRSeCoV and we experimented this system on a few video lessons annotates on the module 'data structure'. The obtained results show the feasibility and benefits of using ontologies to search by the semantic content in pedagogical video segments.
However, it is important to note that our approach is far from being finished and that it has to evolve in the near future.
To improve the research relevance, we think to use semantic inference in the search of content. The results (explicit assertions) returned by the conceptual search may be supplemented by implicit assertions derived or inferred from the knowledge base by exploiting the semantic relations between concepts (e.g. transitivity, similarity, etc…)
We suggest also extending the ontological model, by integrating knowledge about the profile of learners to guide our system to the adaptation of video segments based on their profiles.

Acknowledgments

G. Merzougui would like to thank a lot both M. Moumni and A. Behloul for discussions, comments and suggestions that have greatly enriched the work.

References

[1] V. Psyché, Olavo, J. Bourdeau, Apport de l'ingénierie ontologique au environnements de formation à distance, International Review sticef : science et technologie de l'information et de la communication pour l'Education et la formation, Vol 10, 2003.
[2] M. Charhad, Modèles de Documents Vidéo basés sur le Formalisme des Graphes Conceptuels pour l'Indexation et la Recherche par le Contenu Sémantique, Ph.D, Thesis, University Joseph Fourier, Grenoble, France, 2005.
[3] Dublin Core Metadata Initiative, Available at: http://dublincore.org/documents/dcmi-terms/
[4] ISO/IEC, Overview of the MPEG-7 Standard (version 8), ISO/IEC JTC1/SC29/WG11/N4980, Klagenfurt, July 2002.
[5] R. Troncy, Nouveaux outils et documents audiovisuels: les innovations du web sémantique, 392 Documentaliste – science de l'information, Vol. 42, n°6, 2005.
[6] N. Hernandez, Ontologie de domaine pour la mosélisation du contexte en recherche d'information, Ph.D. Theses, University Paul Sabatier of Toulouse, France, 2006.
[7] R. Troncy, Formalisme des connaissonces documentaires et des connaissances conceptuelles à l'aide d'ontologie: application à la description de documents audiovisuels, Ph.D. Theses, University Joseph-Fourier, Grenoble, 2004.
[8] A. Isaac, R. Troncy, Conception et utilisation d'ontologies pour l'indexation de documents Audiovisuel, Ph. D. These, École doctorale Concepts et Langages. University Paris IV – Sorbonne, 2005.
[9] C. Desmoulins, M. Grandbastien, Des ontologies pour indexer des documents techniques pour la formation professionnelle, Porceesing of the conférence Ingénierie des connaissances. Toulouse (Centre pour l'UNESCO),2000.
[10] A. Benayache, construction d'une mémoire organisationnelle de formation et évaluation dans un contexte elearning : le projet MEMORAe, Ph. D. Theses, l'UTC. 2005.
[11] M. Buffa, S. Dehors, C. Faron-Zucker, P. Sander, Towards a Corporate Semantic Web Approach in Designing Learning Systems, workshop conference AIED Review of the TRIAL Solution Project, 2005.
[12] A. Bouzeghoub, B. Defude, J. Duitama, C. Lecocq, Un modèle de description sémantique de ressources pédagogiques basé sur une ontologie de domaine» revue sticef 'Sciences et Technologies de l´Information et de la Communication pour l´Éducation et la Formation' Vol. 12, 2005.
[13] A. Hammache, R. Ahmed-Ouamer, Un système de recherche d'information pour le e-learning, Revue Document numérique 1279-5127 - VOL 11/1-2 - 2008 - pp.85-105.
[14] A. Behaz, M. Djoudi, Contribution de génération d'un hypermédia d'enseignement adaptatif à base d'ontologies, 3es Journées Francophones sur les Ontologies JFO Poitiers France, 3-4 Décembre 2009.
[15] A. Dong, H. Li, B. Wang, Ontology-driven annotation and Access of Educational Video Data in E-learning, in E-learning Experiences and Future, Edited by: Safeeullah

Soomro, Publisher: InTech, (pp. 305-326, April 2010, ISBN 978-953-307-092-6).

[16] A. Carbonaro, Ontology-based Video Retrieval in a Semantic-based Learning Environment, Journal of e-Learning and Knowledge Society. Vol. 4, n. 3, September 2008 (pp. 203 - 212).

[17] SMIL : Synchronized Multimedia Integration Langage, Available at: http://www.w3.org/AudioVideo/

[18] OWL Web Ontology Language Overview, W3C Recommendation 10 February 2004. http://www.w3.org/TR/owl-features/ . visited on date 17 mars 2011.

[19] G. Salton, M. Mcgill, Introduction to Modern Information Retrieval», McGraw-Hill, 1983.

[20] J. Martinet, Un modèle vectoriel relationnel de recherche d'information adapté aux images, Ph.D. Theses, Unversity of Joseph Fourier – Grenoble I, 2004.[21] H. Zargayouna, Indexation sémantique de documents XML, Ph.D, Theses, University Paris XI Orsay, 2005.

Merzougui Ghalia received a Master in Computer Science from the University of Batna, Algeria, in 2004. She is currently a Professor at the University of Batna, Algeria.
She is a member of (Adaptive Hypermedia in E-learning) research group. She is currently pursuing his doctoral thesis research on the management of multimedia educational content. Her current research interest is in E-Learning, system of information retrieval, ontology, semantic web, authoring and multimedia teaching resource. His teaching interests include computer architecture, software engineering and object-oriented programming, ontology and information retrieval.

Djoudi Mahieddine received a PhD in Computer Science from the University of Nancy, France, in 1991. He is currently an Associate Professor at the University of Poitiers, France.
He is a member of SIC (Signal, Images and Communications) Research laboratory. He is also a member of IRMA E-learning research group. His PhD thesis research was in Continuous Speech Recognition. His current research interest is in E-Learning, Mobile Learning, Computer Supported Cooperative Work and Information Literacy. His teaching interests include Programming, Data Bases, Artificial Intelligence and Information & Communication Technology. He started and is involved in many research projects which include many researchers from different Algerian universities.

Behaz Amel received a Master in Computer Science from the University of Batna, Algeria, in 2004. She is currently a Professor at the University of Batna, Algeria.
She is a member of (Adaptive Hypermedia in E-learning) research group. She is currently pursuing his doctoral thesis research on the modeling of an adaptive educational hypermedia system. Her current research interest is in E-Learning, Knowledge Engineering, Semantic Web, Ontology, and Learner Modeling. Her teaching interests include Programming, Data Bases, and Web Technology.

Diagnosis of Fish Diseases Using Artificial Neural Networks

J.N.S. Lopes[1], A.N.A. Gonçalves[2], R.Y. Fujimoto[1,3] and J.C.C. Carvalho[3]

**[1] Programa de Pós-graduação em Biologia Ambiental, Universidade Federal do Pará
Bragança, Pará, Brazil**

**[2] Programa de Pós-graduação em Genética e Biologia Molecular – com Ênfase em Bioinformática, Universidade Federal do Pará
Belém, Pará, Brazil**

**[3] Docente da Faculdade de Engenharia de Pesca, Universidade Federal do Pará
Bragança, Pará, Brazil**

Abstract

Artificial neural networks (ANNs) are computational intelligence techniques, which are used in many applications, such as disease diagnosis. The objective of this study was to evaluate two artificial neural networks created for the diagnosis of diseases in fish caused by protozoa and bacteria. As a classification system, ANNs are an important tool for decision-making in disease diagnosis. A back-propagation feed-forward was selected, with two layers, sigmoid and linear activation functions, and the Levenberg -Marquardat algorithm, for the training of the ANNs. The results of the application of these neural networks for the diagnosis of fish diseases based on test cases indicated a 97% success rate for the classification of both bacterial and protozoan diseases.

Keywords: *Artificial Neural Networks, Fish Disease Diagnosis, Feed-forward back-propagation network, Artificial Intelligence, and Decision Support Systems.*

1. Introduction

Artificial Neural Networks (ANNs) are adaptive models inspired by the organization of neurons in the human brain and learning of novel patterns from an initial detail [1]. These networks have been used successfully in areas ranging from engineering to medicine [2], [3], [4], [5], [6], [7]. In the medical field, ANNs are used to assist the specialist in the analysis, diagnosis and treatment of diseases. Most medical applications of artificial neural networks are classification problems, in other words, the task is based on the classification of measured parameters, to assign the patient to a small set of classes, which are the diseases [8], [9]. The back-propagation model with two layers (hidden and output), with the sigmoid and linear activation functions, has been used for the solution of some of these problems [6], [9].

In contrast with the field of human medicine, few studies based on neural networks have been developed for the diagnosis of diseases in fish. This is partly because fish diseases are complex phenomena, the diagnosis of which demands considerable expertise, but also because infected fish tend to die quickly without adequate treatment [10]. Economically, the most important diseases include those caused by bacteria and protozoa. These diseases are especially difficult to identify because their clinical signs are similar, and differences may only arise during the acute or chronic phase, and in many cases, transmission patterns are unknown [11], [12]. Given this, there is a clear need for the development of new and more effective approaches to the diagnosis of bacterial and protozoan diseases in these animals.

Zeldis and Prescott [13] discussed problems and solutions for the development of a program for the diagnosis of diseases in fishes. They reviewed the different techniques employed by the experts in the field, emphasizing the considerable difficulties of diagnosing fish diseases, but concluded that the use of artificial neural networks would not be feasible due to the lack of an adequate database of fish diseases.

However, while no central agency store these data, a number of university laboratories have accumulated a large quantity of data, which permits a more systematic evaluation of the phenomenon. In the present study, data from such a source are used to evaluate the applicability of ANNs to the diagnosis of fish diseases caused by bacteria and protozoa. The aim is to provide a reliable system for the rapid and accurate diagnosis of diseases in fishes.

2. Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANNs) are computational systems that simulate biological neural networks, which can also be defined as a specific type of parallel processing system, based on distributional or connectionist methods [14]. Their internal structure can also be modified in accordance with a specific function [1]. The structure of a network of this type is characterized by a number of interconnected elements (neurons) that learn by modifying themselves. As in nature, the function of the network is determined by the connections between the elements [2].

In this configuration of neural networks, a subset of processing elements can be added to the network (layers). This configuration is referred to as a neural network multi-layer perceptron (MLP-ANN). This MLP-ANN is widely- used in applications such as approximation functions, feature extraction, optimization, classification and ease-of- use [15].

In this configuration, the first layer is the input layer and the last, the output layer, between which there may be one or more extra layers, known as hidden layers (see Figure 1). Within the context of a given learning algorithm, this configuration enables neural networks to achieve a specific function, as well as allowing adjustments in the value of the connections (weights) between elements [9], reducing the mean square error.

The back-propagation approach and its variants are widely used as learning algorithms in neural networks. The procedure is based on the calculation of the gradient vector error, with the error gradually decreasing until all the expected results are displayed [2].

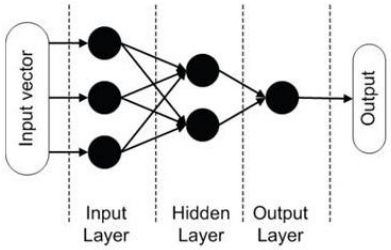


Figure 1 Structure of the multi-layer neural network perceptron.

SOURCE: Krenker, Bester and Kos, 2011 [16].

Artificial neural networks can be divided into two categories - supervised and the unsupervised – based on the learning process. In supervised learning, inputs and outputs are presented to the network, which will adopt the patterns that provide the desired outputs [17]. In order to ensure that the output (system response) achieves a satisfactory result, the neural network adjusts the relative weights of the connections, using an interactive process. Following unsupervised learning, the network develops its own representation of the input stimuli in order to calculate the weights of acceptable connections until finding the answer to the problem. This type of network creates a map of self-organization, which has only inputs and no known responses. An ANN thus becomes a powerful and versatile tool, due to its considerable capacity for learning and, theoretically, that it can provide continuous mapping of any database with arbitrary accuracy [9].

3. The Proposed Method Based on Neural Network

Two back-propagation feed-forward neural networks were constructed, one for bacterial and the other for protozoan diseases. The neural networks were derived from data sets provided by the Ichthyoparasitology and Fisheries Laboratory at the Federal University of Pará (Brazil), which provided information on the clinical signs of diseased fishes and their diagnosis.

The data of the network were divided into inputs and outputs. The input data were the clinical signs, with the presence of signs being scored as 1 (present) or 0 (absent). The output data for each disease group is the diagnosis of the disease.

The network for the diagnosis of bacterial diseases was composed of 43 inputs, 20 neurons in the hidden layer and 12 neurons in the output layer (Figure 2), while that for protozoan diseases had 28 inputs, 22 neurons in the hidden layer and eight neurons in output layer (Figure 3). The structures proposed for the neural networks are presented in Figures 2 e 3.

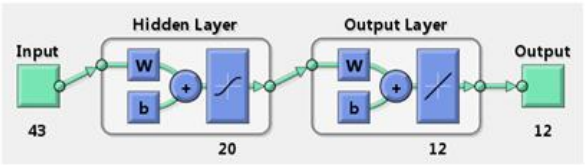


Figure 2 The proposed neural network for the diagnosis of bacterial diseases.

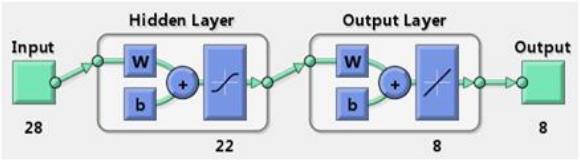


Figure 3 The proposed neural network for the diagnosis of protozoan diseases.

3.1 Feed-forward architecture

This feed-forward neural network model was selected for this project because this approach has been used successfully in other contexts for classification, prediction and troubleshooting. In this network model, information moves in only one direction, always forward from the input nodes, spreading to the hidden nodes and the on to the output nodes, where the output is compared with the desired value, resulting in an error for each element of the output [17].

In this system of feed-forward neural network, the hidden "neurons" are able to learn data patterns during the training phase and are then able to map the relationship between input/output pairs. In the hidden layer, each neuron uses a transfer function to process the data it receives from the input layer and then transfers this processed information on to the neurons of the output layer. The output of the hidden layer can be represented by the following function:

$$YNx1 = f(WNxM\ XM, 1 + bN, 1) \quad (1)$$

where Y is a vector containing the output of each neuron (N) in a layer, W is a matrix containing the weights of each input (M) for all the neurons, X is a vector containing the inputs, b is a vector containing the bias, and f (.) is the activation function [18].

In these types of network, data input and output are automatically divided into training, validation, and test sets. The training data are used for network learning. Training upholds the parameters set in the network structure, and then a set of validation data is used to minimize overfitting. These validation data are used to check the increase in accuracy in comparison with the training data, and are not shown in the final network. If accuracy increases during training, but then validation remains constant or decreases, network training is terminated.

3.2 Fish Disease Diagnosis Data (Protozoan diseases)

A database provided by a specialist in fish diseases was transcribed into the form used for the construction, validation, and testing of the network. The data were then analyzed for the diagnosis of potential diseases caused by protozoa pathogens.

Thirty records of diagnosed diseases caused by protozoa were used, based on the set of clinical signs shown in table 1. Of the 30 samples in the data set, 80% were used to train the neural network, while the remaining 20% were used in the test network.

3.3 Fish Disease Diagnosis Data (Bacterial diseases)

The data on the diseases caused by bacteria were also converted into a form appropriate for analysis in the neural network. Thirty-one records were analyzed based on the clinical signs outlined in table 2. As for the previous procedure, 80% of the samples were used to train the neural network, while the other 20% were used in the test network.

3.4 Performance Evaluation

The tool used for the construction, running and evaluation of the proposed neural networks was Matlab Toolbox 7.10. In both networks, the feed-forward model with sigmoid activation function in the hidden layer and a linear output layer was adopted because it is the most frequently-used procedure for function fitting (or nonlinear regression) problems. The Levenberg-Manquardt training algorithm was also used for the back-propagation network. This type of algorithm is faster for standard and feed-forward networks, and performs better for function fitting (nonlinear regression) than for pattern recognition problems. The network produced in this study can be expected to successfully diagnose eight types of disease caused by protozoa, and 12 by bacteria (Table 3).

4. Results of the Experiment

The result obtained from the artificial neural network approach to the diagnosis of diseases, based on reported clinical signs, demonstrated that the network was able to learn the patterns corresponding to the clinical signs of specific fish diseases. The networks were also subjected to the respective test sets (unknown cases), which again produced satisfactory results, as described below (tables 3 and 4).

4.1 Artificial Neural network 1: Protozoan diseases

The network classified 97% of the cases in the protozoan test set. The validation vectors used to stop the training network at the point set by the training algorithm are shown in figure 4. Validation ceased when the GRADIENT performance decreased, the performance adaptive variable (MU) was reduced, and the validation performance (VAL FAIL) increased.

The best performance validation score (0.01088) was recorded at time 4 (figure 5). The mean square error (MSE) is the mean square of the differences between actual and desired outputs. Lower values indicate better performance, and zero is equal to no error. The validation and test curves were very similar. The percentage accuracy in the sample simulation of the feed-

forward back-propagation network was 97%. Overall MSE was 5.44087e-3 and regression (R) was 9.83867e.-1.

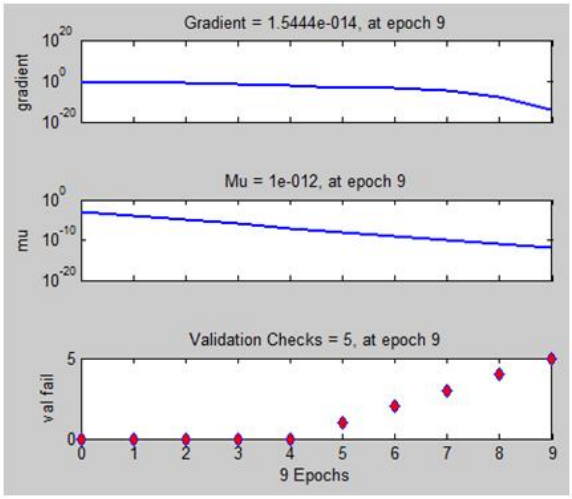


Figure 4: Training state values

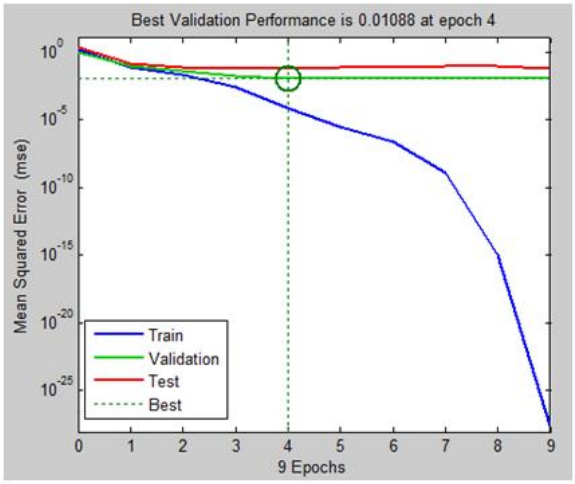


Figure 5: Network error values plot.

Table 3: The Mean Square Error (MSE) and regression values (R) for training, validation and testing.

	MSE	R
Training	7.08966e-5	9.99771e-1
Validation	1.08804e-2	9.99501e-1
Testing	6.44062e-2	9.97507e-1

4.2 Artificial Neural network 2: Bacterial diseases

The second network also classified 97% of the cases in the bacterial test set. The validation vectors used to stop the training network at the point set by training algorithm are shown in Figure 6. Once again, validation ceased when the GRADIENT performance decreased, the performance adaptive variable (MU) was reduced, and the validation performance (VAL FAIL) increased. The best performance validation score (0.056193) was recorded at epoch 7 (Figure 7). The percentage accuracy in the sample simulation of the feed-forward back-propagation network 97%, MSE was 2.28988e-2 and R was 9.54099e-1.

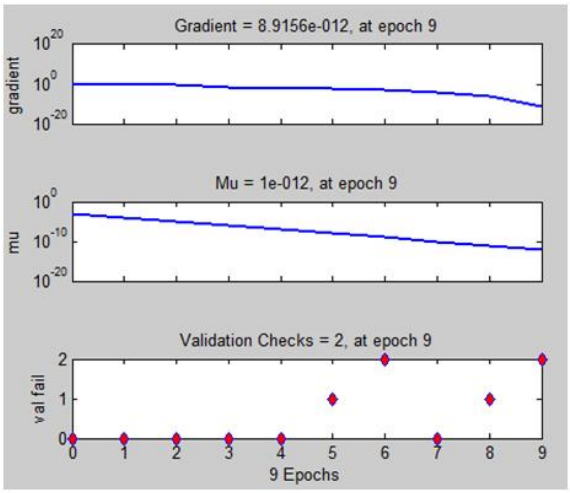


Figure 6: Training state values

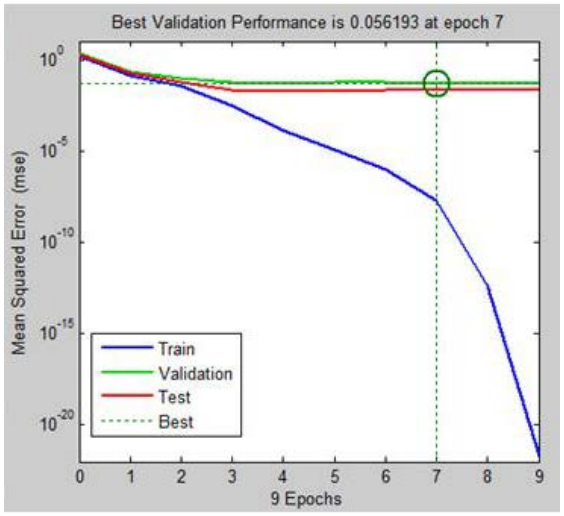


Figure 7: Network error values plot.

Table 4: The Mean Square Error (MSE) and regression values (R) for training, validation and testing.

	MSE	R
Training	1.92955e-8	9.99999e-1
Validation	5.61926e-2	9.14124e-1
Testing	7.65403e-3	9.65968e-1

5. Conclusions and Future Work

This article presents the construction and testing of two feed-forward back-propagation neural networks for the diagnosis of protozoan and bacterial diseases in fishes. The artificial neural networks had satisfactory outcomes for both data sets. The results indicate that artificial neural networks provide a viable approach for the diagnosis of diseases in fish, and may be further enhanced to aid in the treatment of these diseases, as well as the diagnosis of diseases caused by other vectors, such as parasites or fungi, and well as disorders related to environmental changes.

Acknowledgments

"The author J. N. S. Lopes thanks Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES-PROCAD) for financial support".

References

[1] E. Grossi, "Artificial Neural Networks and Predictive Medicine: a Revolutionary Paradigm Shift" - Artificial Neural Networks - Methodological Advances and Biomedical Applications, India: INTECHWEB.ORG, 2011.
[2] G. Caocci, R. Baccoli and G. L. Nasa, "The Usefulness of Artificial Neural Networks in Predicting the Outcome of Hematopoietic Stem Cell Transplantation" - Artificial Neural Networks - Methodological Advances and Biomedical Applications, India: INTECHWEB.ORG, 2011.
[3] G. Zini, "Artificial intelligence in hematology". Hematology, Vol. 10, No 5, 2005, pp. 393-400.
[4] M. Carter, Minds and computers: An introduction to the philosophy of artificial intelligence, Edinburgh: Edinburgh University Press, ISBN 9780748620999, 2007.
[5] M. Suka, S. Oeda, T. Ichimura, K. Yoshida, and J. Takezawa, "Neural Networks Applied to Medical Data for Prediction of Patient Outcome", Trends in Intelligent Systems and Computer Engineering. O. Castillo, L. Xu, and Sio-Iong Ao. Vol. 6, Springer. ISBN 978-0-387-74934-1, 2008.
[6] Q. K. Ai-Shayea, and I. S. H. Bahia, "Urinay System Diseases Diagnosis Using Artificial Neural Networks", International Journal of Computer Science and Network Security (IJCSNS), vol. 10, Nº 7, 2010, pp. 118-122.
[7] R. Linder, I. R. König, C. Weimar, H. C. Diener, S. J. Pöppl, A. Ziegler, "Two models for outcome prediction - a comparison of logistic regression and neural networks", Methods of Information in Medicine, Vol. 45, No. 5, 2006, pp. 536-540.
[8] R. Dybowski, and V. Gant, Clinical Applications of Artificial Neural Networks, Cambridge University Press, 2007.
[9] Q. K. Ai-Shayea, "Artificial Neural Networks in Medical Diagnosis", International Journal of Computer Science Issues (IJCSI), Vol. 8, No 2, 2011, pp. 150-154.
[10] D. Li, Z. Fu and Y. Duan, "Fish-Expert: a web-based expert system for fish diseases diagnosis", Expert Systems with Applications, Vol. 23, 2002, pp. 311-320.
[11] G.C. Pavanelli, J.C. Eiras and R.M. Takemoto, Doenças de Peixes – profilaxia, diagnóstico e tratamento, Maringá: Eduem, 2008.
[12] M.P. Georgiadis, I.A. Gardner, R.P. Hedrick, "The role of epidemiology in the prevention, diagnosis, and control of infectious diseases of fish", Preventive Veterinary Medicine, Vol. 48, 2001, pp.287-302.
[13] D. Zeldis, and S. Prescott, "Fish disease diagnosis program – problems and some solutions", Aquacultural Engineering, Vol. 23, 2000, pp. 3-11.
[14] A. A. Hopgood, Intelligent systems for engineers and scientists, Florida: CRC Press, 2000.
[15] C. M. Bishop, Pattern Recognition and Machine Learning, Cambridge: Springer, 2006.
[16] A. Krenker, J. Bester and A. Kos, "Introduction to the Artificial Neural Networks" - Artificial Neural Networks - Methodological Advances and Biomedical Applications, India: INTECHWEB.ORG, 2011.
[17] S. Sumathi and S. Paneerselvam, Computational intelligence paradigms: theory & applications using MATLAB, CRC Press, 2010.
[18] J. A. Freeman, and D. M. Skapura, Neural networks: algorithms, applications and programming techniques, Addison-Wesley, 1991.

J. N.S. Lopes received her undergraduate degree in Information Systems from the Amazon Higher Education Institute (IESAM), Brazil, in 2009, and will receive her masters degree in Environmental Biology from the Federal University of Pará (UFPA), Brazil, in 2011. Her current research interests include neural networks and fuzzy logic.

A. N. A. Gonçalves, received his undergraduate degree in Computer Engineering from IESAM in 2009 and will receive his masters in Genetic and Molecular Biology with emphasis on Bioinformatics, from UFPA in 2012. He is interested in Artificial Intelligence, programming languages, and databases and algorithms for Bioinformatics.

R. Y. Fujimoto received his undergraduate degree (1998) and masters (2001) in Animal Sciences, and doctorate (2004) in aquaculture from the Julio de Mesquita Filho State University of Sao Paulo. He is currently an adjunct professor at UFPA, where he coordinates the fish farming and ichthyoparasitology laboratory. He has research experience in fish diseases.

J. C. C. Carvalho, received his undergraduate degree (1996), masters (1999) and doctorate (2006) in Electrical Engineering from UFPA. He is currently associate professor of Information Systems I at the UFPA - Bragança campus.

Table 1: Clinical Signs variables used to analyze the data set of protozoan diseases.

Clinical sign of disease	
Nº	**Diagnostic Variable**
1	Abscess {Yes, no}
2	Anorexia {Yes, no}
3	Apathy {Yes, no}
4	Ascites {Yes, no}
5	Cotton-like appearance {Yes, no}
6	Gills with excess mucus {Yes, no}
7	Gills with blood {Yes, no}
8	Blindness {Yes, no}
9	Pale coloration {Yes, no}
10	Dark coloration {Yes, no}
11	Dyspnea {Yes, no}
12	Swimming disorders {Yes, no}
13	Exophthalmos {Yes, no}
14	Injuries to the body {Yes, no}
15	Differentiated feces {Yes, no}
16	Hypertrophy of organs {Yes, no}
17	Bleeding in external organs{Yes, no}
18	Organs with lesions {Yes, no}
19	Ulcerative lesions {Yes, no}
20	White blemishes{Yes, no}
21	Disjointed movements {Yes, no}
22	Fins destroyed {Yes, no}
23	Fins closed {Yes, no}
24	Nodules {Yes, no}
25	White spots {Yes, no}
26	Skin mucus{Yes, no}
27	Rash{Yes, no}
28	Abnormal tegument {Yes, no}

Table 2: Clinical Signs variables used to analyze the data set of bacterial diseases.

Clinical sign of disease	
Nº	**Diagnostic Variable**
1	Abscess {Yes, no}
2	Anorexia {Yes, no}
3	Apathy {Yes, no}
4	Ascites {Yes, no}
5	Swollen anus {Yes, no}
6	Hemorrhagic anus {Yes, no}
7	Hemorrhagic areas{Yes, no}
8	Gills affected {Yes, no}
9	Gills pale {Yes, no}
10	Blindness {Yes, no}
11	Red coloration {Yes, no}
12	Pale coloration {Yes, no}
13	Dark coloration {Yes, no}
14	Abnormal growth {Yes, no}
15	Dyspnea {Yes, no}
16	Fin disorders {Yes, no}
17	Edema {Yes, no}
18	Abnormal scales {Yes, no}
19	Exophthalmos {Yes, no}
20	Furuncle {Yes, no}
21	Hypertrophy in organs {Yes, no}
22	Bleeding {Yes, no}
23	Bleeding in the eyes {Yes, no}
24	Bleeding in the external organs {Yes, no}
25	Bleeding in the internal organs {Yes, no}
26	Minor hemorrhage {Yes, no}
27	White lesions {Yes, no}
28	Dark lesions {Yes, no}
29	Hemorrhagic lesions {Yes, no}
30	Lesions in organs {Yes, no}
31	Minor lesions {Yes, no}
32	Ulcerative lesions {Yes, no}
33	White blemishes{Yes, no}
34	Intense bruising {Yes, no}
35	Membrane surrounding the organs {Yes, no}
36	Fins destroyed {Yes, no}
37	Fins closed {Yes, no}
38	Nodules {Yes, no}
39	Red spots {Yes, no}
40	Delay of sexual maturation{Yes, no}
41	Ulcers {Yes, no}
42	Abnormal tegument {Yes, no}
43	Disease spreading in hours {Yes, no}

Table 3: Types of disease diagnosed for the two groups.

Group	Type of disease
Bacterial	Mycobacteriosis
Bacterial	*Streptococcus* infection
Bacterial	Peduncle disease
Bacterial	"Spinal column" disease
Bacterial	Bacterial gill disease
Bacterial	Septicemia provoked by *Edwardsiella*
Bacterial	Red mouth disease
Bacterial	Furunculosis
Bacterial	Septicemia caused by mobile *Aeromonas*
Bacterial	Septicemia caused by *Pseudomonas*
Bacterial	Bacterial kidney disease
Bacterial	Pseudo-renal disease
Protozoan	Velvet disease
Protozoan	Ichthyobodosis
Protozoan	Disease caused by rhizopods (amoebae)
Protozoan	Disease caused by flagelates (*Hexamita* spp.)
Protozoan	Disease caused by ciliates I (*Trichodina* spp.)
Protozoan	Disease caused by ciliates II (*Chilodonella* spp.)
Protozoan	Disease caused by ciliates III (*Sessilina*)
Protozoan	White spot disease

A Review of Burst Scheduling Algorithm in WDM Optical Burst Switching Network

R.P.Adgaonkar [1] and S.N.Sharma [2]

[1] Department of Computer Engineering, G.S. Mandal's Marathwada Institute of Technology (M.I.T.) Aurangabad, M.S., India

[2] Department of Computer Engineering, G.S. Mandal's Marathwada Institute of Technology (M.I.T.) Aurangabad, M.S., India

Abstract

Optical Burst Switching (OBS) has proved to be an efficient paradigm for supporting IP-over-WDM networks. The growth of a variety of applications which transmit voice, data, video and multimedia, has necessitated the need to provide Quality of Service (QoS) over OBS networks. One of the key factors in OBS is the scheduling algorithm that is used in the switches to allocate the incoming bursts to a wavelength. Since the arrival of bursts is dynamic, it is highly desirable that the scheduling is done as quickly as possible. In this paper, a survey of various existing burst scheduling algorithm that provide QoS and reduce burst dropping probability is presented and compare different algorithm.

Keywords: WDM, Optical Burst Switching Network, LAUC

1. Introduction

A WDM technology has the enormous amount of bandwidth available in fiber cable. In WDM system, each carries multiple communication channels and each channel operating on different wavelength. Such an optical transmission system has a potential capacity to provide Tera bytes of bandwidth on a single fiber. WDM technology has the capability to provide the bandwidth for the increase in the huge on traffic demand of various applications like audio, video and multimedia, which needs the QoS over the network [1].

The currently existing switching techniques can be broadly classified into optical circuit switching (OCS), optical packet switching (OPS) and OBS techniques [1], [2]. In OCS, an end-to-end optical light path is setup using a dedicated wavelength on each link from source to destination to avoid optical to electronic (O/E/O) conversion at each intermediate nodes. Once the light path is setup, data remain in optical domain throughout transmission of data. OCS is relatively easy to implement but main drawback of OCS is circuit setup time and improper holding time of resources like bandwidth. On other hand, no circuit setup is required in OPS but packet header need to be processed in the electronic domain on hop-by-hop basis. Due to which data payload must wait in optical buffers like fiber delay lines (FDLs), which is very complex and challenging task in high speed optical networks. To do this task, OPS require optical buffers, O/E/O converters and synchronizers. The new switching technology, which combines the merits of coarse gained OCS and fined gained OPS was proposed and called as OBS [1], [2], [3], [4].

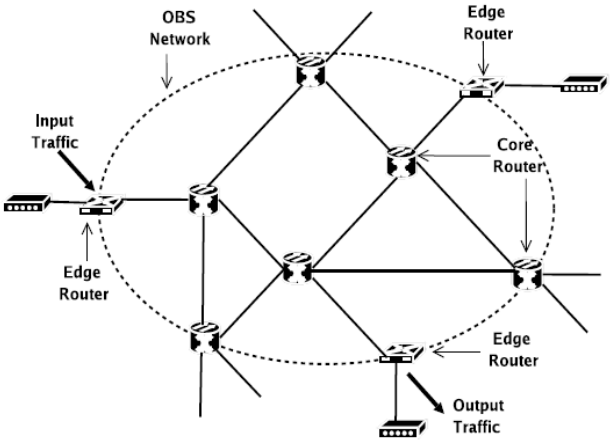


Fig 1: OBS Network Model

In OBS network model, as shown in the Fig. 1 [5], there are two types of routers, edge and core router, which are connected by WDM links. Various type of client's data with same destination are aggregated at the edge router in a data burst. The data could be IP/SONET/SDH/ATM cell or combination of all packet type.In OBS, edge router is responsible for burst assembly/ disassembly, scheduling of burst, transmission of burst, deciding the offset time, generation of burst control packet (CP) functions. Core router will forward the burst to its destination node [6],

[7]. In OBS, a burst consist of header and payload called data burst. A burst header is called as CP. Typically; CP contains information about burst size and burst arrival time. The CP and payload are send separately on different channels called as control and data channel respectively as shown in Fig. 2 [8]. The burst is preceded in time by a CP, which is send on separate control wavelength. The preceded time is called as "offset time".

After a burst is generated, the burst is buffered in the queue at edge router for an offset time before being transmitted to give its CP enough time to reserve network resource along its route. During offset time, packets belonging to that queue may continue to arrive. These extra packets are dropped [9].

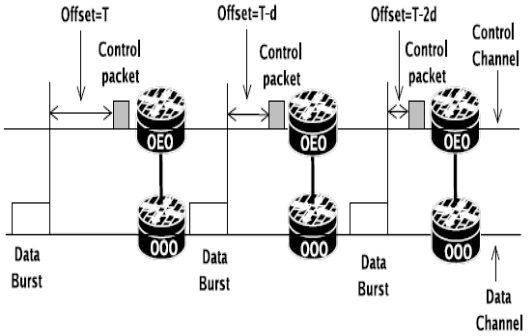


Fig 2: Separate Transmission of data and control signals.

At each intermediate node, CP undergoes O/E/O conversion to get it processed electronically. The time taken for processing a CP is called as the "processing time" [9], [10], [11]. Depending upon CP information wavelength is reserved for the incoming burst for that duration by core router [4], [6],[7], [8].

Basically, there are three different assembly schemes, namely threshold-based, timer-based and hybrid-based [9], [10].

In a timer-based scheme, a timer is started to initialize the assembly process. A burst containing all the packets in the buffer is generated when the timer exceeds the burst assembly period [9]. While in a threshold-based scheme, a burst is created and sends into the OBS network when the total size of the packets in the queue reaches threshold value [9].

Hybrid assembly scheme is the combination of both threshold-based and timer-based assembly scheme [9]. In the hybrid assembly scheme, a burst can be sending out when either the burst length exceeds the desirable threshold value or the timer expires.

In OBS network, different wavelength reservation schemes are used for reserving the wavelength. One is called as Tell-And-Wait (TAW). In TAW, when source has the burst to send, it first reserve the wavelength along the route by sending "request" message. If the wavelength is granted by intermediate nodes along its route, a positive acknowledgment (PACK) message returns to source from the destination; otherwise negative acknowledgment (NACK) is received at source [4], [12],[13], [14].

Second scheme is called Tell-And-Go (TAG), in which two reservation schemes has been proposed. They are Just-Enough-Time (JET) and Just-In-Time (JIT). In JET, reservation is made by using CP information. The reservations made for the duration of data burst. The resources are reserved and released implicitly. In JIT, the resources are reserved as soon as CP is received and hold resources until burst departure time. The resources are released explicitly by sending another control message and which results in bad resource utilization. Due to this the wavelength holding time to that node is larger than burst transmission time [4], [12], [13], [15].

2. Burst Scheduling Algorithm

Another important factor which affects the network traffic is scheduling algorithms used to schedule burst. Arrival of bursts at OBS node is dynamic. Scheduling technique must schedule arrival burst on the available wavelengths for the entire duration of burst transmission. Scheduling technique must schedule burst efficiently and quickly. Scheduling algorithm should be able to process the CP fast enough before the burst arrives to the node. It should also be able to find proper void for an incoming burst to increase channel bandwidth utilization. Following are proposed burst scheduling algorithms in the literature.

2.1 Latest Available Unused Channel (LAUC) Algorithm [16],[17].

In LAUC, burst scheduling is done by selecting the latest available unscheduled data channel for each arriving data burst. In this algorithm, a scheduler keeps track of horizon for each channel. Horizon is the time after which no reservation has been made on that channel. LAUC searches the wavelength by using horizon information on each channel. The scheduler assigns each arriving new burst to the data channel with minimum void formed by that burst on data channel. For example, in Fig.3, wavelength $C2$ and $C3$ is unscheduled at the arrival time t of the new burst. Wavelength $C3$ will be selected for the new burst because the generated void (t-$t3$) on wavelength $C3$ will be smaller than the void (t-$t2$) that would have been created if wavelength $C2$ was selected.

LAUC algorithm is simple and has a good performance in terms of its execution time. However, it results in low bandwidth utilization and a high burst loss rate.

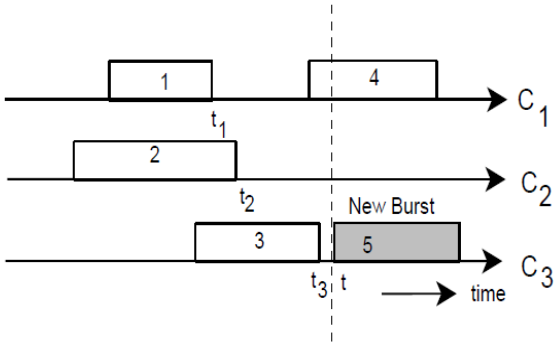


Fig 3: Illustration of LAUC data scheduling algorithm.

2.2 Latest Available Unused Channel with Void Filling (LAUC-VF) Algorithm [16],[17].

In LAUC, the voids are created between two data burst assignment on the same data channel. This is termed as unused channel capacity. LAUC-VF is variant of LAUC. In this algorithm, a scheduler keeps track of horizon and voids for each channel. LAUC-VF maintains start and end time of void for each data channel. LAUCVF searches for the void such way that newly formed void is very small compared to other voids. An example of LAUC-VF algorithm is illustrated in Fig. 4. New data burst with duration *L* arrives at time *t* to the optical switch, the scheduler first finds the outgoing wavelengths that are available for the time period (*t*, *t*+*L*). Wavelengths *C*1, *C*2 and *C*5 are available for the coming data burst. Wavelengths *C*2 is chosen to carry the new data burst because the void that will be produced between the bursts and coming data burst is the minimum void.

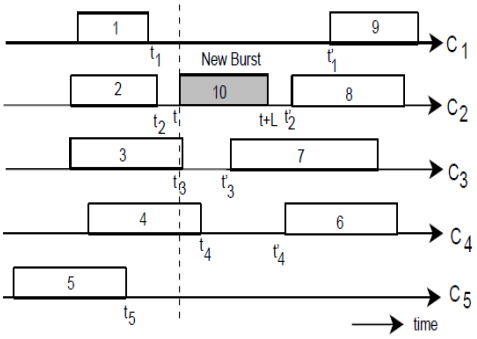


Fig 4: Illustration of LAUC-VF data scheduling algorithm

Implementation of LAUC-VF has a much longer execution time than the LAUC scheduling algorithm, especially when the number of voids is significantly larger. However, it result in high bandwidth utilization and a low burst loss rate.

2.3 Best-Fit (BF) Algorithm [16].

In BF, a scheduler keeps track of horizon and void for each channel. It also maintain start time and end time of void for each data channel. Scheduler tries to search for a void such way that newly created void is the smallest void before and after scheduled burst. An example of Min-EV algorithm is illustrated in Fig. 5.

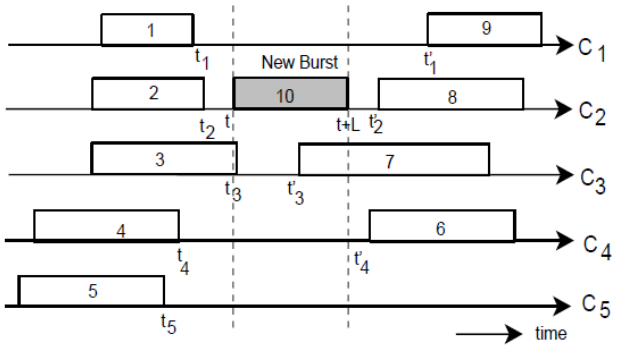


Fig 5: Illustration of BF data scheduling algorithm

New data burst with duration *L* arrives at time *t* to the optical switch, the scheduler first finds the outgoing wavelengths that are available for the time period (*t*, *t*+*L*). Wavelengths *C*1, *C*2, *C*4 and *C*5 are available for the coming data burst. Wavelength *C*2 is chosen to carry the new data burst because the starting and ending void that will be produced between the bursts and coming data burst is the minimum void. Implementation of BF has a much longer execution time than the LAUC scheduling algorithm, especially when the number of voids is significantly larger. Also it achieves a loss rate which is at least as low as LAUC-VF, but can run much faster. However, it results in high bandwidth utilization and a low burst loss rate.

2.4 Minimum Starting Void (Min-SV) Algorithm [16], [17].

In Min-SV, a scheduler keeps track of horizon and void for each channel. It also maintains start and end time of void for each data channel. Scheduler tries to search for a void such way that newly created void is the smallest void after scheduled burst. An example of Min-SV algorithm is illustrated in Fig. 6. New data burst with duration *L* arrives at time *t* to the optical switch, the scheduler first finds the outgoing wavelengths that are available for the time period (*t*, *t*+*L*).Wavelengths *C*1, *C*2 and *C*5 are available for the coming data burst. Wavelength *C*2 is chosen to carry the new data burst because the starting void that will be produced between the burst and coming data burst is the minimum void.

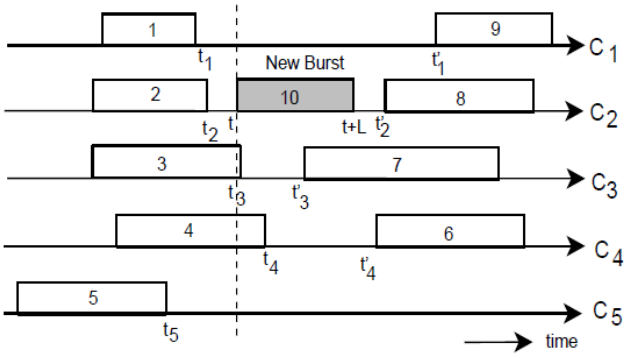


Fig 6: Illustration of MIN-SV data scheduling algorithm.

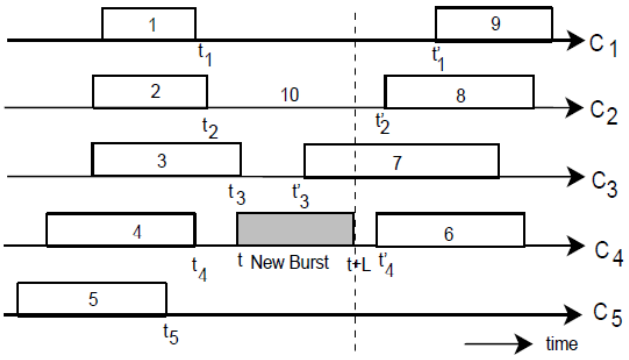


Fig 7: Illustration of MIN-EV data scheduling algorithm.

Implementation of Min-SV has a much longer execution time than the LAUC scheduling algorithm, especially when the number of voids is significantly larger. Also it achieves a loss rate which is at least as low as LAUCVF, but can run much faster. However, it results in high bandwidth utilization and a low burst loss rate.

2.5 Minimum Ending Void (Min-EV) Algorithm [16],[17].

In Min-EV, a scheduler keeps track of horizon and void for each channel. It also maintain start and end time of void for each data channel. Scheduler tries to search for a void such that newly created void is the smallest void before scheduled burst. An example of Min-EV algorithm is illustrated in Fig. 7. New data burst with duration *L* arrives at time *t* to the optical switch, the scheduler first finds the outgoing wavelengths that are available for the time period (*t*, *t*+*L*). Wavelengths *C*1, *C*2, *C*4 and *C*5 are available for the coming data burst. Wavelength *C*4 is chosen to carry the new data burst because the ending void that will be produced between the bursts and coming data burst is the minimum void. Implementation of Min-EV has a much longer execution time than the LAUC scheduling algorithm, especially when the number of voids is significantly larger. Also it achieves a loss rate which is at least as low as LAUCVF, but can run much faster. However, it result in high bandwidth utilization and a low burst loss rate.

BF, Min-SV and Min-EV algorithms are the variant of LAUC-VF algorithm. All the void filling scheduling algorithm yields better bandwidth utilization and burst loss rate than LAUC algorithm. But all the void filling scheduling algorithm has a longer execution time than LAUC algorithm.

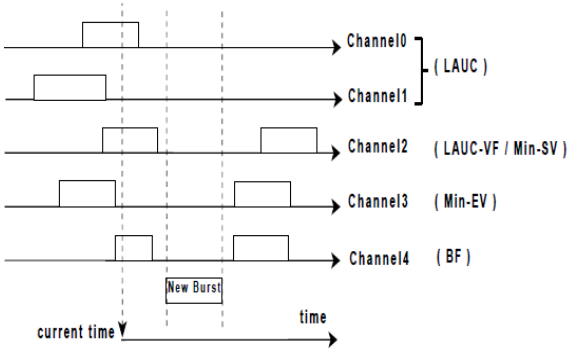


Fig 8: An example showing how a new burst is scheduled by using different scheduling algorithm.

Table I shows the comparison of different scheduling algorithm [16].

Table 1: Comparison of Different Scheduling Algorithm.

Scheduling Algorithms	Time Complexity	State Information	Bandwidth Utilization
LAUC	O(W)	$Horizon_i$	Low
LAUC-VF	O(Wlogm)	$S_{i,j}E_{i,j}$	High
BF	O(Wlogm)	$S_{i,j}E_{i,j}$	High
Min-SV	O(logm)	$S_{i,j}E_{i,j}$	High
Min-EV	O(logm)	$S_{i,j}E_{i,j}$	High

Table I summarizes the above discussion using the following notations:

- W : Number of wavelengths at each output port.\
- m : Maximum number of data bursts(or reservations) on all channels.
- *Horizoni* : Horizon of the *ith* data channel.
- *Si;j* and *Ei;j* : Starting and ending time of *jth* reservation on channel i.

3. Conclusions

OBS provides a cost-effective solution for switching in the next-generation optical Internet. Various Internet applications such as multimedia, voice-over-IP, ecommerce and web conferencing have different resource requirements and differ in how much they are willing to pay for the services. In this paper, a survey of burst rescheduling algorithms in OBS is presented along with advantages and disadvantages of this algorithm.

References

[1] B. Mukharjee *Optical WDM Networks*, Springer Publication 2006.

[2] Tzvetelina Battestilli and Harry Perros, "An Introduction to Optical Burst Switching", *IEEE Optical Communication ,Aug 2003*

[3] C. Qiao and M. Yoo, "Optical Burst Switching: A New Paradigm for An Optical Internet", *Journal of High Speed Network*, vol. 8, pp. 69-84, 1999.

[4] Takuji Tachibana and Soji Kasahara, "Performance analysis of timer-based burst assembly with slotted scheduling for optical burst switching network", *PERFORMANCE EVALUATION An International Journal*, vol. 63, pp. 1016-1031, 2006.

[5] Yuhua Chen and Pramod K. Verma, "Secure Optical Burst Switching: Framework and Research Directions", *IEEE Communication Magazine*, 2008.

[6] Tzvetelina Battestilli, "Optical Burst Switching: A Survey", *Technical Report*, NC State University, Computer Science Department, July 2002..

[7] Y. Chen, C. Qiao and X. Yu, "An Optical Burst Switching: ANew Area in Optical Networking Research", *IEEE Netwoks*, vol.18, pp. 16-23, 2005..

[8] B. Praveen, J. Praveen and C. Siva Ram Murty, "A Survey of differentiated QoS schemes in optical burst switched networks", *SCIENCE DIRECT Optical Switching and Networking*, vol. 3, pp. 134-142, July 2006.

[9] Jason P. Jue and Vinod V. Vokkarane, *Optical Burst Switching Networks*, Springer Publication, 2005.

[10] Wang Ruyan, Wu Dapeng and Guo Fang, "Data Burst Statictics and Performance Analysis of Optical Burst Switching Networks with Self-Similar Traffic", *IEEE Computer Society*, 2007.

[11] Burak Kantarci, Sema F. Oktug and Tulin Atmaca, "Performance of OBS techniques under self-similar traffic based on various burst assembly techniques", *Computer Communications*, vol. 30, pp. 315-325, 2007.

[12] Karamitsos Ioannis and Varthis Evagelos, "A Survey of Resrvation Schemes for OBS", University of Aegean, Department of information and Communication Systems.

[13] Noureddine Boudriga, "Optical burst switching protocols for supporting QoS and adaptive routing", *ELSEVIER Computer Communications*, vol. 26, pp. 1804-1812, 2003

[14] G. Mohan, K. Akash and M. Ashish, "Efficient techniques for improved QoS performance in WDM optical burst switched networks", *ELSEVIER Computer Communications*, vol. 28, pp. 754-764, 2005.

[15] Konstantinos Chistodoullopoulos, Emmanouel Varvarigos and Kyriakos Vlachos, "A new burst assembly scheme based on average packet delay and its performance for TCP traffic", *Optical Switching and Networking*, vol. 4, pp. 200-212, 2007.

[16] Jinhui Xu., Chunming Qiao, Jikai Li and Guang Xu, "Efficient Channel Scheduling Algorithms in Optical Switched Networks using Geometric Technique", *IEEE Journal on selected areas in Communication*, vol. 22, No. 9, November 2004.

[17] Jikai Li and Chunming Qiao, "Schedule burst proactively for optical burst switching networks", *ELSEVIER Computer Networks*, vol. 44, pp. 617-629, 2004.

First Author Sanjay N. Sharma has Received B.E. (Computer Engineering) degree in 1992. M.B.A. in (Marketing Management) in 1995.Currently pursuing M.E (Computer Engineering) from Aurangabad. He has more than 18 years of experience in teaching. Currently working as Assistant Professor in Computer Engineering at Thakur College of Engineering, Kandivali East, Mumbai, INDIA.His areas of interest are Image Processing and Computer Networks.

Second Author Dr. R.P. Adgaonkar has received B.E. (Hons.) from Government College of Engineering, Aurangabad in 1968, M.Tech from IIT Bombay in 1970,and Ph.D.in 1980.

Real-Time Projection Shadow with Respect to Sun's Position in Virtual Environments

Hoshang Kolivand[1] , Azam Amirshakarami[2] and Mohd Shahrizal Sunar[3]

[1,2,3] ViCubelab, Department of Computer Graphics and Multimedia, Faculty Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310, Skudai Johor, Malaysia

Abstract

This paper proposes a real-time software for outdoor rendering to control the shadow's position with effect of sun's position. The position of sun plays an important rule for outdoor games. Calculation of sun's position, as a result, position and length of shadows require a lot of attention and preciseness. Julian dating is used to calculate the sun's position in the virtual dome. In addition, of computer graphics, building design is another field that this paper contributes on it. To create shadow, projection shadow is proposed. By calculating the sun's position in the specific date, time and location on the earth, shadow is generated. Length and angle of shadow are two parameters measured for building design and both of them are calculated in this real-time application. Therefore, it can be used for teachers to teach some part of physics about earth orbit and it can be used in building design and commercial games in virtual reality systems.

Keywords: *real-time shadow, sun's position, sun light, projection shadow*

1. Introduction

The principle calculations of the sun's positions have been very well known for a long time. The ancient Egyptians were able many years ago to calculate the sun's position so. By digging a large hole inside one of the pyramids, just once a year, on the king's birthday, the sun could shine on the grave of their king.

To create a realistic environment, shadow is the most important effects used to reveal information about the distance between objects in the scene. It is the major factor of 3-D graphics for virtual environment but unfortunately, it is difficult to implement in virtual environments, especially in real-time games. In computer games, shadows give the gamers feelings that trigger the sense that they are playing in the real world, resulting in maximum pleasure. Games which lack shadow are not seen as attractive by gamers, especially since gamers' have had a taste of virtual games, and their imagination now requests more and more realistic situations when they are watching cartoons or playing games.

There are some different shadow techniques, like drawing a dark shape similar to the occluder on a plane. Although it is not precise, it is frequently used, especially in old computer games and some parts of advertisement animation. Another simple method to create real-time shadow is projection shadow algorithm [1] that is still widely used in game programming. In this method, shadow can be created just on horizontal and vertical plants at a time, but to create shadow on two adjacent horizontal and vertical planes need more calculation [2]. To have a shadow on arbitrary objects stencil buffer is appropriate.

In computer games, shadow can reveal real distance between objects in virtual environment and give the gamer's maximum feeling. A computer game without shadow cannot be very attractive even for indulged users when they play games or watch cartoons.

In 1999, Preetham et al. approached an analytic model in rendering the sky. The image that they generated is attractive [3]. They present an inexpensive analytic sky model from Perez et al. (Perez model) that approximates full spectrum daylight for various atmospheric conditions [4].

In 2008, [5], worked on sky color with effect of sun's position. They used Julian dating and Perez model to create sky color. In 2010, Sami M. Halawani et al. published a paper entitled "Interaction between sunlight and sky color with effect of sun's position" [6]. They used Julian dating to control the position of sun. In 2008, Ibrahim Reda et al. introduced precise formulas for Julian day and used it for solar radiation [7].

2. Methods

For outdoor rendering the sun's position, sky color and shadows are most important effects. In this paper, the sun's

position and shadow are combined. Projection shadow is high-speed method to create shadow on flat surfaces. The Julian date is one of the accurate techniques to determine the position of the sun.

2.1 Hard Shadows

Projection shadow technique is one of the simple methods to generate a real-time shadow. The most important advantage of projection shadows is high-speed rendering. The most prominent drawback of projection shadow is the fact that it needs a huge calculation to have shadow on arbitrary objects. The main idea of this method is to draw a projection of each occluder's pixels on the shadow receiver along the ray that is started from the light source up to the plane [2].

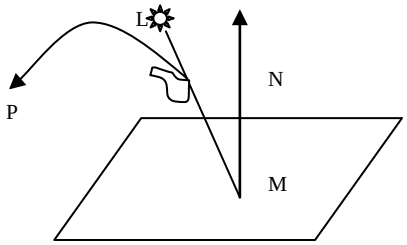


Fig. 1. The scenario of shadow projection

L is light source
P is a pixel of occluder
M is projection of P
N: Normal vector of ground.

$$(\vec{x} - \vec{E})\vec{n} = 0$$

$$\vec{x} = \vec{L} + \lambda(\vec{P} - \vec{L})$$

$$M = L + \frac{E}{NP - D}(P - L)$$

$$\text{Shadow matrix} = \begin{pmatrix} LxNx + E & LxNy & LxNz & -ELx - DLx \\ LyNx & LyNy + E & LyNz & -ELy - DLy \\ LzNx & LzNy & LzNz + E & -ELz - DLz \\ Nx & Ny & Nz & -D \end{pmatrix}$$

By using projection matrix for each pixel of occluder, projection shadow will be appearing on the plane.

2.2 Dome Modeling

Latitude is a distance from north to south of the equator. Longitude is the angular distance from east to west of the prime meridian of the Earth. Longitude is 180 degree from east to west. Each 15 degree represent one hour of each time. For example, if you can travel towards west 15 degree per hour, you can turn off your time and turn on your time up on arrival without having any change in time. The earth spins around the sun in specific orbit once year.

Dome is like a hemisphere in which the view point is located inside it. To create a hemisphere using mathematical formulas the best formula is:

$$f(\theta, \varphi) = cos^2\theta cos^2\varphi + sin^2\theta + cos^2\varphi sin^2\theta - r^2 \quad (1)$$

Where θ is the zenith and φ is the azimuth and
$0 \leq \theta \leq \frac{\pi}{2}$
$0 \leq \varphi \leq 2\pi$
This ranges is needed for a dome on the above of observer and the rest of sphere is not needed [11].
Before creating shadows in virtual environment, the sun's position must be determined; and this will be described in the next section

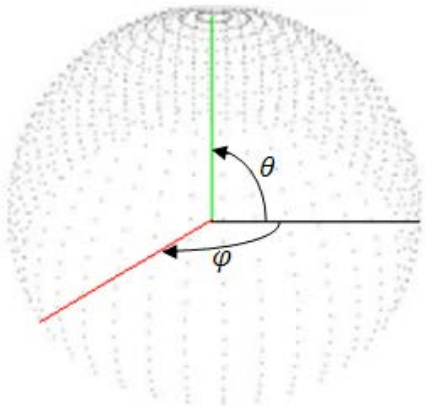


Fig. 2. The zenithal and azimuthal angles on the hemisphere that θ is latitude and φ is longitude

2.3 Sun's Position

The principle calculation of the sun's position is well known long time ago and some exact data are needed. The ancient Egyptians were able in many years to calculate the sun's position so, with digging a large hole inside one of the pyramids, just once a year, when it is also the birthday of the king; the sun could shine on the grave of their king [8].

The earth's oriented North - South line is not exactly perpendicular to the orbit. It has about 23.5° deviation. The diversion of earth during a turn in the orbit around the sun maintains. When earth is located on the right side of the sun, the southern hemisphere, due to the slight

deviation (23.5°), more direct radiation from the sun receives. About six months later, when the earth goes to the other side of the sun, this radiation to the northern hemisphere is vertical.
Longitude and latitude are two most important necessary aspects to calculate the sun's position. The other information that is required is Greenwich Mean Time (GMT). To determine the position of the sun in the created dome, zenith and azimuth are enough.

2.4 Calculation of Sun's Position Using Julian Date

To calculate position of the sun, zenith and azimuth are enough. To have zenith and azimuth, location, longitude, latitude, date and time are needed [9]. Zenith is the angle that indicates the amount of sunrise while the azimuth is the angle that indicates the amount angle that sun turns around the earth.
In 1983, Iqbal [10] proposed a formula to calculate the sun's position and in 1999, Preetham et al.[4] improved it. It is a common formula to calculate the position of the sun in physics.

$$t = t_s + 0.17sin\left(\frac{4\pi(j-80)}{373}\right) - 0.129sin\left(\frac{2\pi(j-8)}{355}\right) + 12\frac{SM-L}{\pi} \quad (2)$$

where
t: Solar time
t_s: Standard time
J: Julian date
SM: Standard meridian
L: Longitude

$$\delta = 0.4093sin\frac{2\pi(j-81)}{368} \quad (3)$$

The solar declination is calculated as the following formula:
δ: Solar declination
The time is calculated in decimal hours and degrees in radians.
Finally zenith and azimuth can be calculated as follows:

$$\theta_s = \frac{\pi}{2} - \sin^{-1}(sinl\ sin\delta - cosl\ cos\delta cos\frac{\pi t}{12}) \quad (4)$$

$$\varphi_s = \mathrm{Tan}^{-1}\left(\frac{-cos\delta sin\frac{\pi t}{12}}{cosl\ sin\delta - sinl\ cos\delta\ cos\frac{\pi t}{12}}\right) \quad (5)$$

where
θ_s : Solar zenith
Φ_s : Solar azimuth

l : Latitude

With calculation of zenith (θ_s) and azimuth (φ_s) the sun's position is obvious.

2.5 Effect of Sun's Position on Shadows

As the earth moves in its orbit, sun is moved from earth. Shadow's length depends on the position of the sun relative to the view situation. When a part of earth is tilted away from the sun, the sun's position is lower in the sky and shadows are long. On the other hand, when a part of earth is tilted towards the sun, position of the sun is highest in the sky and shadows are short. Longer shadows in a day appear at the sunrise and sunset; and the short shadows appear at the noon. The tilted of the earth axis is a reason that each part of earth can see more or less sun in a day. Different season is a result of different length of days and nights.

3. Result and Discussion

Sun's position and length of shadows in real-time computer games can make a game realistic as much as possible. To keep real position and length of shadows in a virtual environment a substantial amount of precision is needed. Solar energy is free and a blessing of God for us, optimized usage of this blessing needs a mastermind. On the other hand, in building design and architecture, possible recognition of which direction is best to build a building in specific location. In cold places, building needs to stay in a situation that shadows lie in the back of the building but on the contrary, in warm places building should be located in the direction that shadows lie in front of the building.
Figure 3, 4 and 5 show the direction of sun shine and shadow at 7:40, 10:59 (b) and 16:30 respectively, in UTM university on latitude 1.28 and longitude 103.45 in 22 April 2011. The viewer location can be changed by changing the longitude and latitude. Date and time are also changeable. One of the facilities of the software is rendering automatically during a daytime.

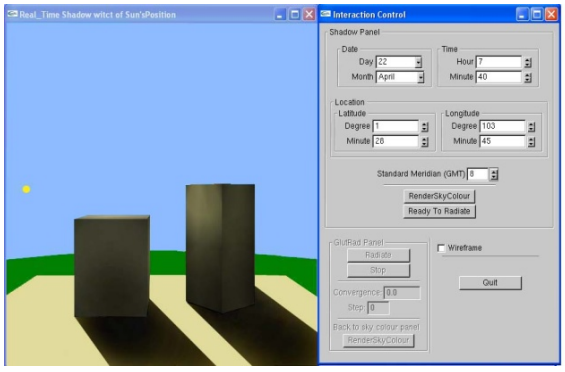


Fig. 3. Result of application at 7:40 in UTM (April 22, 2011)

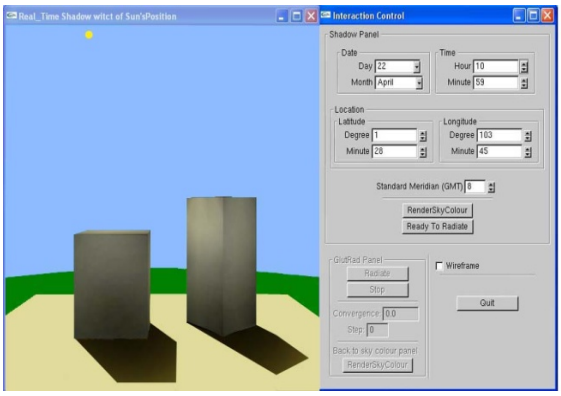


Fig. 4. Result of application at 10:59 in UTM (April 22, 2011)

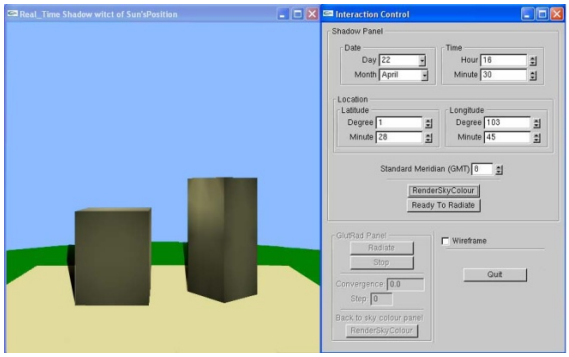


Fig. 5. Result of application at 16:30 in UTM (April 22, 2011)

4. Future work

In this study, we have focused on projection shadow on flat surfaces. Shadow volume and shadow mapping are other techniques to create shadow. To have shadows on arbitrary object, shadow volume using stencil buffer is appropriated. Shadow mapping is more convenient for outdoor rendering because of high-speed rendering but not more than projection shadow. Volume shadow using stencil buffer or shadow mapping combined with sky color can make it as much realistic as possible. Interaction between the sun's position, sunlight, shadow and sky color in the outdoor virtual environment can be more attractive.

5. Conclusion

Since this research targets at shadows and sun's position for game engines and virtual environment, the real-time shadow technique seems appropriate. Subsequently, the hard shadow such as projection shadow was rendered with sun's position to get the precise effect of sun's position in outdoor effects such as shadows. The methodology used in this research combines projection method with sun's position and sky modeling. The first objective of this research is achieved by the use of projection shadow to create hard shadows and to recognize the position of the sun for each viewpoint in specific location, date and time of day.

The proposed application can be used for outdoor games without any hesitation about the position of the sun and location of shadow for each object. Other contribution of this application is for building designer to find the best direction to build a building to save the solar energy. Expert proposed method can also be used to display the sun's position and to describe the amount of shadow changes in some high schools.

ACKNOWLEDGEMENTS

This research was supported by UTMVicubeLab at Department of Computer Graphics and Multimedia, Faculty of Computer Science and Information System, Universiti Teknologi Malaysia. Special thanks to Universiti Teknologi Malaysia (UTM) Vot. 00J44 Research University Grant Scheme (RUGS) for providing financial support of this research.

References

[1] F.Crow, “Shadow algorithms for computer graphics”, Computer Graphics, Vol. 11, No.2,pp. 242-247, 1977.

[2] H. Kolivand, M.S. Sunar ," Real-Time Shadow Using A Combination OF Stencil And The Z-Buffer”, The International Journal of Multimedia & Its Applications , Vol.3, No.3,pp. 27-38 , 2011

[3] R. Perez, R. Seals, Michalsky, J., “All-weather model for sky luminance distribution - preliminary configure and validation”, Solar Energy, Vol. 50, No. 3 (1993).

[4] A.J. Preetham, P Shirley, B. Smith, “A practical analytic model for daylight”, Computer Graphics (Siggraph '99 Proceedings), pp. 91-100, 1999.

[5] A.B.M Azahar, M.S. Sunar, D. Daman, et al. “Survey on real-time crowds simulation, Technologies For E-Learning and Digital Entertainment, Proceedings Vol. 5093 pp. 573-580, 2008

[6] S.M. Halawani,M.S. Sunar, "Interaction Between Sunlight And The Sky Color With 3d Objects In The Outdoor Virtual Environment",2010 Fourth International Conference On Mathematic/Analytical Modeling And Computer Simulation.

[7] I. Reda, A. Andreas, “Solar position algorithm for solar radiation applications”, National Renewable Energy Laboratory (2008).

[8] S. Nawar, A.B. Morcos, J.S. Mikhail, “photoelectric study of the sky brightness along sun’s meridian during the march 29 2006”, Solar Eclipse, New Astronomy, Vol. 12 , pp. 562–568, 2007

[9] T. P. Chang, “The sun’s apparent position and the optimal tilt angle of a solar collector in the northern hemisphere”, Solar Energy, Vol. 83, pp. 1274–1284,2009

[10] M. Iqbal, 1983. “An introduction to solar radiation”, Academic Press. 390.
[11] K.R Nirmal, N. Mishra, 3D Graphical User Interface on personal computer using p5 Data Glove”, International Journal of Computer Science Issues Vol. 8, Issue 5, No 1 pp155-160, 2011

Authors

Hoshang Kolivand received the B.S degree in Computer Science & Mathematic from Islamic Azad University, Iran in 1997, M.S degree in Application Mathematic and computer from Amirkabir University, Iran in 1999. He is now pursuing Ph.D in UTM ViCunelab under the guidance of Dr Mohd Shahrizal Bin Sunar. His research interests include Computer Graphics. Previously he was a lecturer in Shahid Beheshti University Iran. He has published enormous articles in international journals and conferences as well as national journals, conference proceedings and technical papers including article in a book. Hoshang Kolivand is an active reviewer of some conferences and international journals. He has published four books in object-oriented programming and one in mathematics.

Azam Amirshakarami received the Associate Degree in Software Engineering (2002) from Islamic Azad University, Iran; BSc degree (2005) from Shariati University, Iran and MSc in Computer Science majoring in Computer Graphics (2010) from Universiti Teknologi Malaysia . She is Ph.D candidate since 2010 from Universiti Teknologi Malaysia under the supervision of Dr Mohd Shahrizal Bin Sunar. Her major field of study is improving real-time interaction in outdoor augmented reality. Prior to her current position, she managed the IT team at ISESCO regional office in Tehran beside other activities such as software developing and teaching software design and programming in C#, Java, Visual Basic. This year she is serving as a faculty member at a private

Mohd Shahrizal Sunar received the BSc degree in Computer Science majoring in Computer Graphics (1999) from Universiti Teknologi Malaysia and MSc in Computer Graphics and Virtual Environment (2001) from The University of Hull, UK. In 2008, he obtained his PhD from National University of Malaysia. His major field of study is real-time and interactive computer graphics and virtual reality. He is the head of computer graphics and multimedia department, Faculty of Computer Science and Information System, Universiti Teknologi Malaysia since 1999. He had published numerous articles in international as well as national journals, conference proceedings and technical papers including article in magazines. Dr. Shahrizal is an active professional member of ACM SIGGRAPH. He is also a member Malaysian Society of Mathematics and Science.

Designing an Improved Fuzzy Multi Controller

Saeed Barzideh [1], Arash Dana [2], Ahmad Ali Ashrafian[2] Gh.Sajedy Abkenar[3]

[1] Scientific Association of Electrical & Electronic Eng. Islamic Azad University Central Tehran Branch Tehran, Iran

[2] Dept. of Elect. Eng., Central Tehran Branch, Islamic Azad University Tehran, Iran

[3] Scientific Association of Electrical & Electronic Eng. Islamic Azad University Central Tehran Branch Tehran, Iran

Abstract

Fluid level, pressure, temperature and flow control is a very common problem in the industry. Considering the advantages of fuzzy control systems instead of other conventional methods of control, in this paper a new approach using fuzzy logic to control the fluid level of biphasic (liquid and steam) fluid system, based on both temperature and pressure parameters presents. Simulations show that joining "pressure changes" parameter as third input of fuzzy controller, improves the efficiency of our control system and lead to more smooth changes in its output.

Keywords: *Fuzzy controller, Level, Pressure, Temperature.*

1. Introduction

Benefits of Artificial Intelligence (AI) based control systems compared to other classical control methods, has encouraged many interested researchers to study and design such systems. Reviewing and comparing between conventional and intelligent control systems, especially discussion about classical and fuzzy logic methods, in many articles shows the importance of AI in new controllers. Combination of conventional control methods and fuzzy is also an attractive research area for many researchers [1-4].

Artificial Intelligence based Control methods, especially in the case of non-linear systems or when system has many complications, are very efficient, because in these cases the system cannot be addressed simply by equations and mathematical descriptions.

In fuzzy control systems, human knowledge in the form of fuzzy if-then rules is the foundation for decision making in fuzzy inference system (FIS). In a fuzzy control system, which system parameters are crisp numbers, they must change into fuzzy sets by Fuzzifier to be able to react as the fuzzy inference engine inputs. Fuzzy inference engine interprets fuzzy input sets and assesses them with fuzzy if-then rules, finally the results will expressed by fuzzy output sets. Since, the equipments react by crisp inputs, fuzzy output sets have to change into crisp numbers by Defuzzifier. The various part of a fuzzy control system are shown in Fig. 1.

In the following paragraphs, in section two, controller design methods and section three simulation methods are described. Results are discussed in Section four and finally in section five conclusions and future works are expressed.

2. Design Method

In this paper pressure, temperature and pressure-derivation considered as proposed FIS inputs. In practice, information of temperature and pressure comes by sensors as crisp numbers and pressure derivation is obtained by calculating at any time. As shown in Fig. 1, these three inputs, after fuzzification will be given to FIS. Fuzzy inference engine is responsible for decision making according to if-then rules database. Final results will be explains as fuzzy sets. We used Mamdani's[1] method for this decision making. More details are available in [5].

In order to control the fluid level in the boiler, based on two parameters temperature and pressure, usually two separate control mechanisms and two separate outlet valves are considered. In this paper, using fuzzy logic, a control surface will be defined by the percent outlet valve openness, based on both these parameters. This surface is represented in (Fig. 2). To improve the efficiency of this mechanism, first derivation of the pressure -as third FIS input- is jointed in the final decision making. In reference

1. Ebrahim Mamdani [Mam75]

[6] by S. Panich, the similar method is presented for controlling the output level of a tank with pressure variable from 0 to 12 bars and temperature from 0 to 120 ° C. The pressure and temperature are the controller input and percent of output valve openness is the controller output. Fuzzy surface that obtained by Panich method is something similar to Fig. 2 that obtained in our simulations. But, because we have used three inputs, two other fuzzy surfaces are available for analyses: one surface represent controller output depending upon the "pressure" and "pressure derivation" and another surface, related to the "temperature' and "pressure derivation". For instance Fig. 3 shows the percent of outlet valve openness versus "pressure" and "pressure derivation".

However in this method, we increased applied FIS rules up to 75 but it is not so complicated for new control systems to analyze 75 rules instead of 25. New added rules will increase the system flexibility and this will make a smooth and more careful outlet valve operation. We will discuss more about outlet valve operation in various situations and through the system input changes in the fourth section.

Most of the time and depends on system contents and features, changes in temperature could be sense with some delay for obvious reasons. But in the case of pressure changes, instruments usually could sense variations almost as soon as theirs happens. This is the reason that prods us to elect this parameter as third input of fuzzy inference system.
On the other hand considering water thermodynamic properties [7], in our certain range (160 to 230 ° C and pressure 17 to 21 bars) temperature and pressure of biphasic system (water and steam) are also in thermodynamically equilibrium. So regardless of changes in other parameters, reducing or increasing the temperature and pressure should be commensurate.

3. Simulation Method

Mass may not be transferred in or out of a close system boundaries that always contain the same amount of matter whereas heat and work could be exchanged across the boundary of the system. In open systems, matter may flow in and out of the system boundaries. The first law of thermodynamics for open systems states: the increase in the internal energy of a system is equal to the amount of energy added to the system by matter flowing in and by heating, minus the amount lost by matter flowing out and in the form of work done by the system [7].

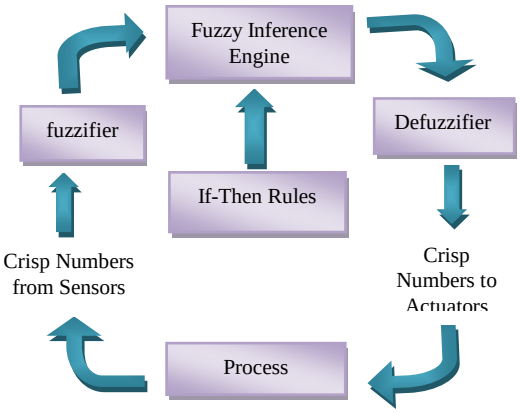


Fig. 1. Various part of fuzzy control system.

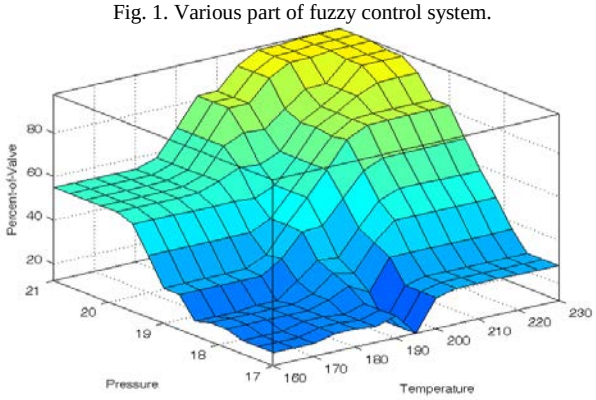


Fig. 2. Output control surface.

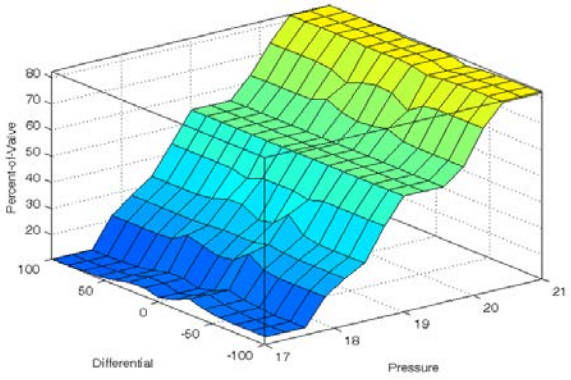


Fig. 3. Percent of outlet valve openness versus pressure and pressure derivation.

In simulation, for better investigations, sometimes we have considered that pressure and temperature changes amongst each other. That is because basically we don't experience a closed system. And for example the boiler input water may affect our system properties.

We have used MATLAB7.5 to design and simulate controller. Temperature has considered in the range from 160 to 230 ° C and pressure could vary between 17 to 21 bars. It is an under pressure tank operation range, that could be applied as a part of super hot steam production chain, in power plants and refineries.

3.1. Fuzzification and Defuzzification Parts

In FIS the Fuzzifier classifies crisp input parameters, temperature and pressure into five fuzzy sets, from lowest value to the highest. All input and output linguistic variables have been normalizing in the range from 0 to 1. Very Low (VL), Low (L), Medium (M), High (H), and Very High (VH) are the linguistic variable terms that could be considered to describe the severity of two input variables temperature and pressure. We have used trapezoidal membership function for the sake of its simplicity and applicability for those parameters. Fuzzy membership functions for temperature and pressure are shown in Fig. 4 and Fig. 5. Using sigmoid curve membership function, pressure derivation could be expressed by three linguistic variables: negative, zero and positive. Fig. 6 depicts membership function for input variable pressure derivation. After calculation, pressure derivation crisp data, has been normalized in the range of -100 to +100. Percent of valve openness is the single system output. Because we are trying to have more softly changes in outlet valve operation, we have used Gaussian membership function in defuzzification part. Being very smooth, Gaussian membership function will provide our favourite results. Strong Close (SC), Close (C), Medium (M), Open (O) and Strong Open (SO) are our five linguistic variables in FIS defuzzification part (Fig. 7).

3.2. Inference Engine and If-Then Rules

The main part of decision making in our algorithm is depended on pressure and temperature parameters changes, and only some percent (in this paper about 15%) of outlet valve operation is considered to affect by the

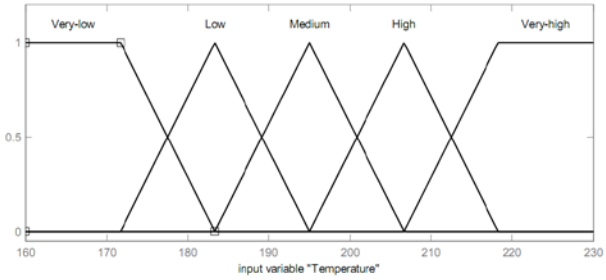


Fig. 4. Membership function for input variable: temperature.

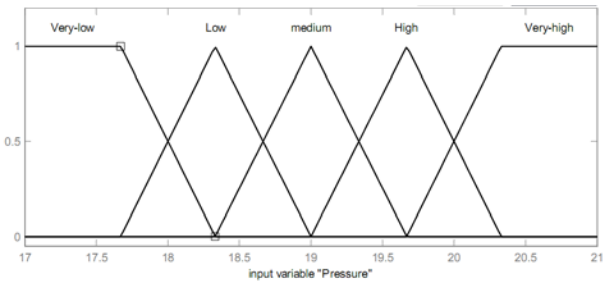


Fig. 5. Membership function for input variable: pressure.

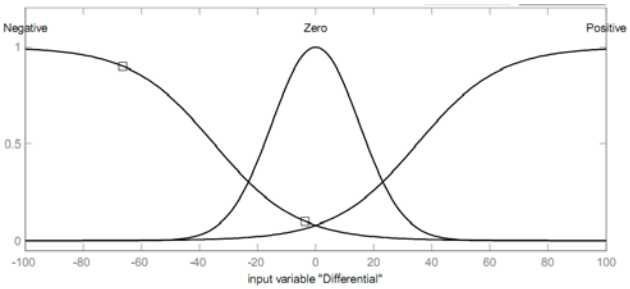


Fig. 6. Membership function for input variable: differential.

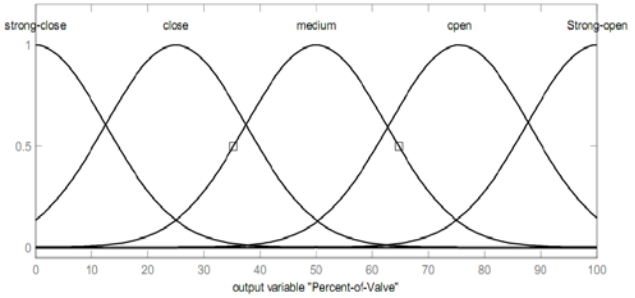


Fig. 7. Membership function for output variable: percent of valve openness.

parameter “temperature-derivation”. This percent can prepare or change by if-then rules and depended on designer practical experience. It is expected that this method make outlet valve operation more smooth and improve its reaction by forecasting changes in system parameters. We will investigate this idea during our simulations. In the simulations that carried out in this paper, 75 fuzzy if-then rules have been used for decision making in FIS. 75 if-then rules constitute a three dimensional rule table. Each rule in our proposed method will participate in decision making with a weight number between 0 and 1, which represent severity of effect of that particular rule. In this way the control surface (Fig. 2) is extremely flexible and under the designer desire control. The fifteen first rules are as bellow:

1. If (Pressure is Very-low) and (Differential is Positive) and (Temperature is Very-low) then (Valve is strong-close)(1)
2. If (Pressure is Low) and (Differential is Positive) and (Temperature is Very-low) then (Valve is strong-close) (1)
3. If (Pressure is medium) and (Differential is Positive) and (Temperature is Very-low) then (Valve is close) (1)
4. If (Pressure is High) and (Differential is Positive) and (Temperature is Very-low) then (Valve is medium) (1)
5. If (Pressure is Very-high) and (Differential is Positive) and (Temperature is Very-low) then (Valve is medium) (1)
6. If (Pressure is Very-low) and (Differential is Zero) and (Temperature is Very-low) then (Valve is strong-close) (0.6)
7. If (Pressure is Low) and (Differential is Zero) and (Temperature is Very-low) then (Valve is strong-close) (0.6)
8. If (Pressure is medium) and (Differential is Zero) and (Temperature is Very-low) then (Valve is close) (0.6)
9. If (Pressure is High) and (Differential is Zero) and (Temperature is Very-low) then (Valve is medium) (0.6)
10. If (Pressure is Very-high) and (Differential is Zero) and (Temperature is Very-low) then (Valve is medium) (0.6)
11. If (Pressure is Very-low) and (Differential is Negative) and (Temperature is Very-low) then (Valve is strong-close) (0.2)
12. If (Pressure is Low) and (Differential is Negative) and (Temperature is Very-low) then (Valve is strong-close) (0.2)
13. If (Pressure is medium) and (Differential is Negative) and (Temperature is Very-low) then (Valve is close) (0.2)
14. If (Pressure is High) and (Differential is Negative) and (Temperature is Very-low) then (Valve is medium) (0.2)
15. If (Pressure is Very-high) and (Differential is Negative) and (Temperature is Very-low) then (Valve is medium) (0.2)

3.3. Defuzzifier Part

FIS classifies its outputs in five fuzzy sets: Strong Close (SC), Close (C), Medium (M), Open (O) and Strong Open (SO). These sets must change into crisp numbers for being applicable in actuators and tools. We used Center of Gravity method to defuzzify FIS outputs. Defuzzifier output, explains percent of valve openness by a crisp number that could be apply to actuators. In other similar cases for example, this number might use to express a buster pump power. Fig. 8 has graphically depicted some part of defuzzification by center of gravity method.

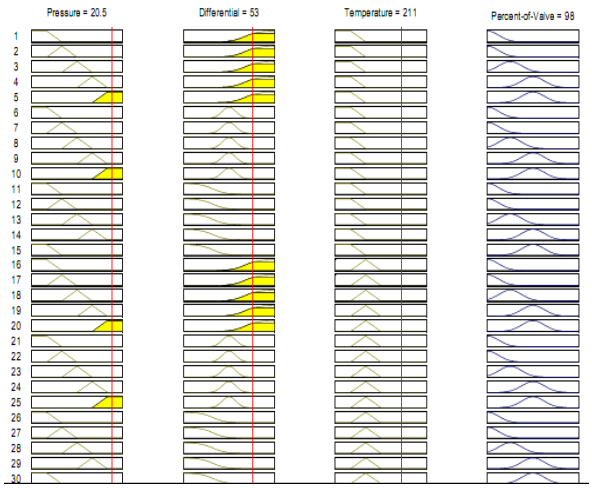


Fig. 8. Defuzzification by center of gravity method.

4. Simulation Analysis

Simulation output results are shown in Fig. 9. The output of fuzzy controller with only two fuzzy inputs designed by Panich is shown in blue. We called this method ***"fuzzy controller"***. And proposed controller output, with three fuzzy inputs is shown in red, and under the title of ***"improved fuzzy controller"***. We applied temperature and pressure to controllers as those inputs. The controller's output will be investigated in different cases and with changes in those parameters. Obviously the controller must response to increasing and decreasing of its inputs by opening and closing some proportionate percent of outlet valve.

According to simulation output results (Fig. 9), at time 15, reducing in the temperature is correctly diagnosed by improved fuzzy controller and its outlet valve is shown appropriate response by closing for about 15 percent. Fuzzy controller doesn't show any suitable response to this situation. At time 30 increasing in pressure on both controllers have been correctly diagnosed. But the slope has more gentle changes in the proposed controller.

At time 35 the pressure has been kept fixed, improved controller was started to opening the outlet valve more softly considering the increment in temperature. After that, increasing in temperature lead to increase the speed of valve opening in the improved controller that shows the prediction capability of improved fuzzy controller. But we can't see remarkable sensitivity in reaction of Panich's fuzzy controller.

Both controllers respond the same near the time 40. After that, improved fuzzy controller has been senses the reduction of the temperature and starts to closing the valve. Fuzzy controller reaction takes place with some delay. At time 45 we have a minimum in temperature. After that time

with increasing in temperature fuzzy controller starts to opening the valve immediately but improved fuzzy controller starts the same action with some delay. Both of these two behaviours of improved fuzzy controller are reasonable, because of atypical minimum (fault) in temperature reporting. (We have a minimum in temperature while pressure has been kept constant near to its maximum). This is a fault in system considering the water and vapour thermo dynamic equivalence carves as mentioned by Gordon J.Van Wylen in [7]. In fact in this case improved fuzzy controller compensates this fault by increasing in temperature, keeping the valve in its close state some more.

Decreasing in pressure while the temperature is constant is also an unusual case in time 70, Because of above mentioned reasons. We discuss this case only for study.

There is a little point to mention too. As we could see in times 5 and 50, sometimes when it is needed to close the outlet valve, fuzzy controller opens it a little before start to closing! Improved fuzzy controller shows it's better reaction in that cases too and have a smooth operation at reminded times.

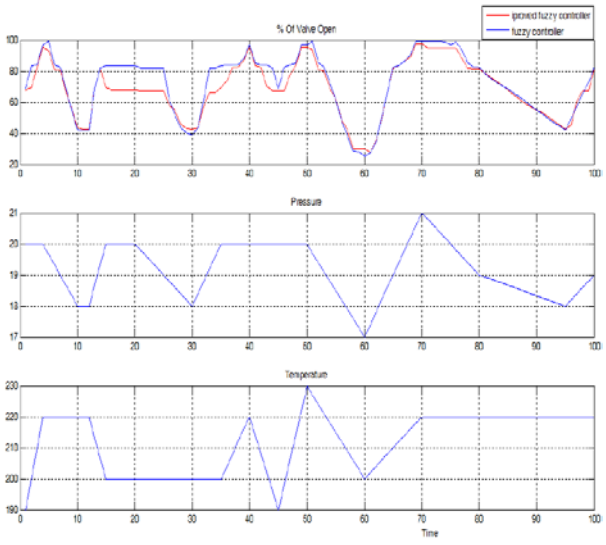


Fig. 9. Simulation output results.

5. Conclusions and Future Works

This is a very important problem in mechanical fluid systems to have smooth changes in fluid flow. Sever changes in flow may cues mechanical damages in system considering water hammer phenomenon. Water hammer (or, more generally, fluid hammer) is a pressure surge or wave resulting when a fluid (usually a liquid but sometimes also a gas) in motion is forced to stop or change direction suddenly (momentum change). Water hammer commonly occurs when a valve is closed suddenly at an end of a pipeline system, and a pressure wave propagates in the pipe. It may also be known as hydraulic shock.

This pressure wave can cause major problems, from noise and vibration to pipe collapse. It is possible to reduce the effects of the water hammer pulses with accumulators and other features, [8].

In this paper a new method based on fuzzy logic was present to controlling the biphasic fluid level according to both temperature and pressure parameters. Industry usually uses two separate mechanisms and two independent control valves to carry out this problem. Simulations show that joining pressure changes, as third input parameter to fuzzy controller improve its behavior in controlling the outlet valve operation. In this paper fuzzy surface for a particular situation was considered. It is possible to optimize if-then rules of this surface for other situation and problems too. In this paper we considered some percent of outlet valve operation relevant to defined third parameter. It caused our controller to have gently operations and also being more flexible and sensitive. Depended on designer experience and problem necessity in similar cases the penetration of the last parameter could be optimized.

References

[1] V. KUMAR, K.P.S. RANA and V. GUPTA "Real-Time Performance Evaluation of a Fuzzy PI + Fuzzy PD Controller for Liquid-Level Process" INTERNATIONAL JOURNAL OF INTELLIGENT CONTROL AND SYSTEMS VOL. 13, NO. 2, pp. 89-96, JUNE 2008.

[2] M. Suresh and G. J. Srinivasan and R. R. Hemamalini," Integrated Fuzzy Logic Based Intelligent Control of Three Tank System" SERBIAN JOURNAL OF ELECTRICAL ENGINEERING Vol. 6, No. 1, pp. 1-14, May 2009,.

[3] E. Natsheh, K. A. Buragga" Comparison between Conventional and Fuzzy Logic PID Controllers for Controlling DC Motors" IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 5, September 2010.

[4] D. Kolokotsa," Comparison of the performance of fuzzy controllers for the management of the indoor environment" Available online at www.sciencedirect.com

[5] Timothy J. Ross,"FUZZY LOGIC WITH ENGINEERING APPLICATIONS" John Wiley & Sons, 2004.

[6] S. Panich," Development of Fuzzy Controller for Water Level in Stream Boiler Tank" Journal of Computer Science 6 (11) pp. 1233-1236, 2010.

[7] Gordon J.Van Wylen and Richard E. Sonntag,"Fundamentals of Classical Thermodynamics (6th Edition)" Jhon Wiley &Sons 2002.

[8] Thorley, ADR, "Fluid Transients in Pipelines", 2nd ed. Professional Engineering Publishing, 2004

Saeed Barzideh received his B.S degree in control & instrumentation in 2007 from the enghlab eslami college Tehran, Iran. Currently, he is developing his electronics M.Sc. degree at Islamic Azad University Central Tehran Branch. His research interests are in the area of analysis and design advanced control

systems based on Artificial Intelligence and also Industrial Data communications and applications like wireless sensor networks.

Arash Dana and **Ahmad Ali Ashrafian,** are academic members of Electrical Engineering Department, Islamic Azad University Central Tehran Branch, Tehran, Iran.

G.h Sajedy Abkenar is a member of Scientific Association of Electrical & Electronic Eng. Islamic Azad University Central Tehran Branch. Tehran, Iran. He is also a student member of IEEE.

Design of a New Model of Multiband Miniature Antenna Near Isotropic

Abdellatif Berkat[1], Noureddine Boukli-Hacene[2]

[1] Telecommunication Laboratory, Faculty of Technology, Abou-Bekr Belkaïd University Tlemcen, 13000, Algeria

[2] Telecommunication Laboratory, Faculty of Technology, Abou-Bekr Belkaïd University Tlemcen, 13000, Algeria

Abstract

In this paper, we propose a new slotted multiband antennas simulated at different frequencies. The insertion of slots in the patch gives a good adapting frequency with various forms on the radiation pattern. The main feature of the proposed antenna is the capability to generate a near isotropic radiation pattern in different frequencies .The design details of the conceived antenna are presented and discussed. Simulations of the different reflection coefficient and radiation pattern are presented. These were carried out using *CST Microwave Studio*. This model has got numerous applications in network sensors, field measurements and electromagnetic compatibility.

Keywords: *Multiband antenna, slot antenna, miniature antenna, near isotropic coverage, circular polarization.*

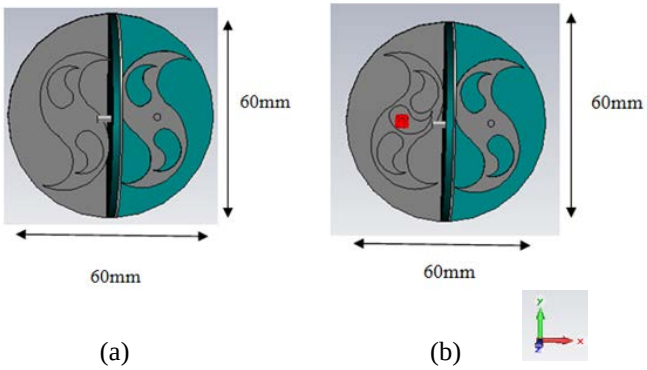
1. Introduction

In recent years the rapid growth of wireless communication technologies leads to the great demands in using a multi-frequencies band on one device, because the systems want to operate at multi-frequencies in some applications such as mobile applications, pico-cell base station applications, and Wireless LAN applications. The main purpose is to reduce number of antennas in the systems. The antenna operating in multi-frequency operation band has been invented and it is called a multiband antenna [1]. However, the transmitted signal is expected to be as stable as possible, whatever the orientation of the communicating objects is. For short distance, low cost, low data rate and low consumption applications, that is to say, when an adaptive solution cannot be envisaged, the most straight-forward strategy is to search for an antenna radiating uniformly in all directions, knowing that an isotropic antenna doesn't exist [2].

The geometry and detailed dimensions and the feeding network as well as the far-field pattern results of the proposed antenna are successively presented below.

2. Antenna structure

In this study, we propose two models of geometry with the main difference is the number of slots in the vertical patch .The antenna structure is depicted in Fig1. Six slotted patches are located along the sides of two intersected cylinders, such as, four patches on the horizontal mode and two patches on the vertical mode. As presented the following figure.

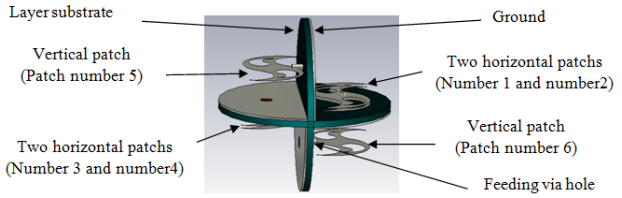


(a) (b)

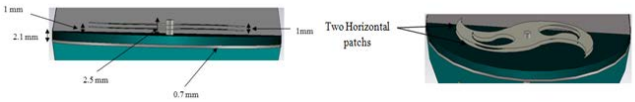
Fig1-(a)Antenna with two slots in the vertical patch,(b) Antenna with tree slots in the vertical patch of the new model

Fig 2(a) illustrate the antenna structure. The green top layer is made of a low permittivity and low-loss substrate, in order to optimize the antenna efficiency and bandwidth, where $\varepsilon_r = 2.33$, $\mu = 1$ and thickness = 2.1 mm. A 0.7 mm thick copper layer is used as a ground plane for the antenna

structure. Fig 2(b) illustrate the horizontal patches structure ,such as, the distance between two patches on each other is 1mm. As presented the Fig 2(b) .



(a)



(b)

Fig 2-(a) Structure of the antenna with six patches, (b) Structure of the horizontal patches.

The six patches are fed with equal amplitudes. S2 and S3 are fed with the same phase of 90°. There is a phase difference of 90° between S1, (S2,S3) and S4. It presents the advantage of greatly reducing the mutual coupling between patches.

Patch number	1 and2	3and 4	5	6
Amplitude relative to patch	1	1	1	1
Phase delay relative to patch	0°	90°	90°	180°

Table1: Amplitude and phase constraints of the antenna.

3. Vertical patch configuration

3.1 Patch with two slots

Fig3 illustrates the elementary vertical patch are 36.85 mm long, 23.79 mm wide .And two slots of 12.21 mm long,9.35 wide . This patch is made of copper. They are fed via holes of 1 mm wide and 2.5mm long. Via holes are connected to the feeding network.

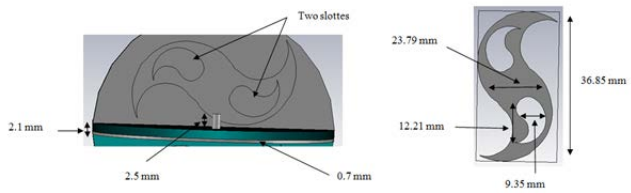


Fig 3: Vertical patch used with two slots

3.1.1 Radiation properties of antenna

In Fig4, the computed return loss of antenna with two slots in vertical patches. The simulated antenna by *CST Microwave Studio* software is well adapted at five resonant frequencies of 4.7 GHz, 5.5GHz, 6GHz, 7.4GHz and 8 GHz. The reflected power reaches the values of -21.74 dB,-14.5 dB,-18.4 dB,-13.5 dB and –12.83 dB at these resonant frequencies respectively.

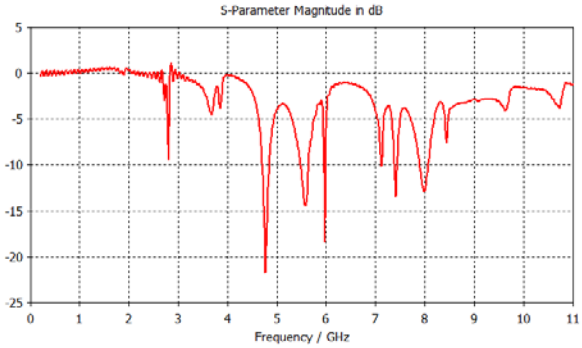


Fig4: Computed return loss of antenna

The main purpose of the antenna is it's near isotropic radiation pattern which allows the communication performances to be uniform between devices whatever are their orientations in several resonant frequencies . The antenna radiation pattern is near isotropic. Fig 5 (a) to Fig 5(o) presents the antenna directivity pattern in two cutting plans at different resonant frequencies.

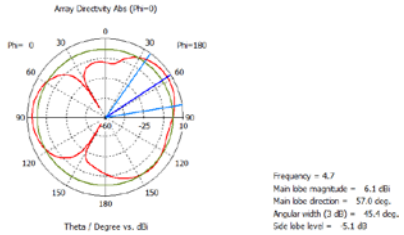


Fig 5(a) : Polar diagrams (Phi=0°) at frequency = 4.7 GHz

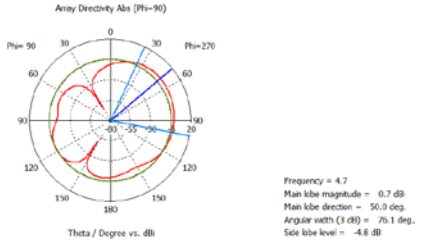


Fig 5(b) : Polar diagrams (Phi=90°) at frequency = 4.7 GHz

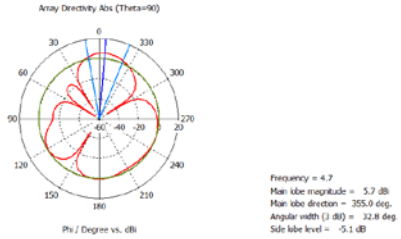


Fig5(c) : Polar diagrams (Theta=90°) at frequency = 4.7 GHz

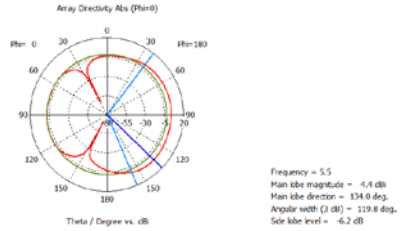


Fig5(d) : Polar diagrams (Phi=0°) at frequency = 5.5 GHz

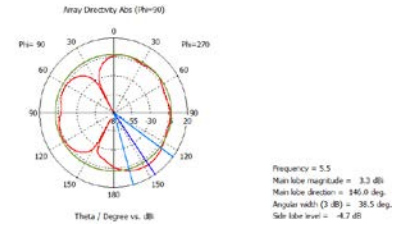


Fig5(e) : Polar diagrams (Phi=90°) at frequency = 5.5 GHz

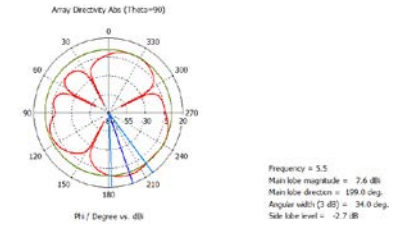


Fig5(f) : Polar diagrams (Theta=90°) at frequency = 5.5 GHz

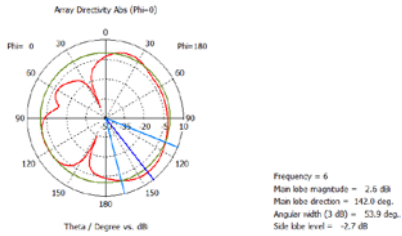


Fig5(g) : Polar diagrams (Phi=0°) at frequency = 6 GHz

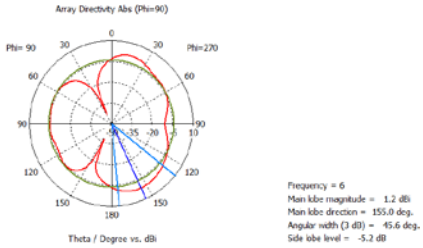


Fig5(h) : Polar diagrams (Phi=90°) at frequency = 6 GHz

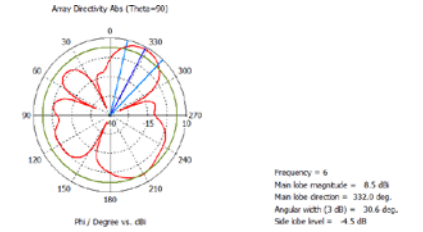


Fig5(i) : Polar diagrams (Theta=90°) at frequency = 6 GHz

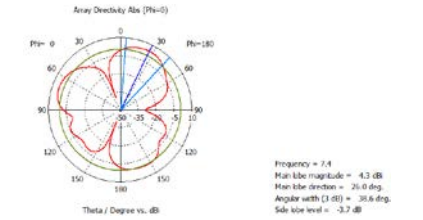


Fig5(j) : Polar diagrams (Phi=0°) at frequency = 7.4 GHz

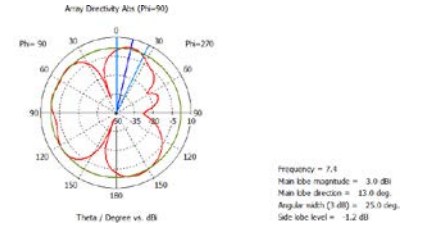


Fig5(k) : Polar diagrams (Phi=90°) at frequency = 7.4 GHz

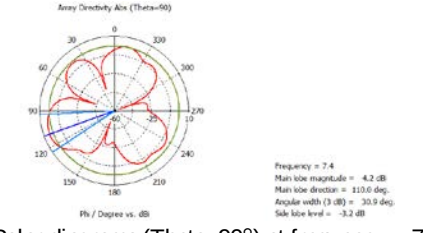


Fig5(l) : Polar diagrams (Theta=90°) at frequency = 7.4 GHz

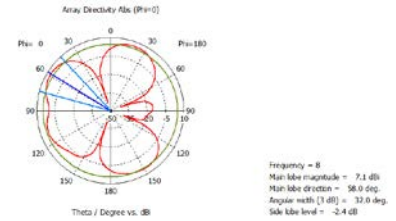


Fig5(m) : Polar diagrams (Phi=0°) at frequency = 8 GHz

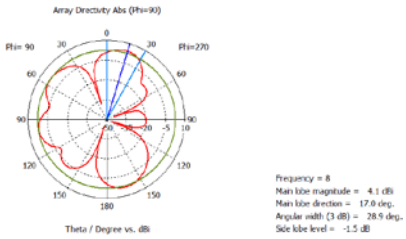


Fig5(n) : Polar diagrams (Phi=90°) at frequency = 8 GHz

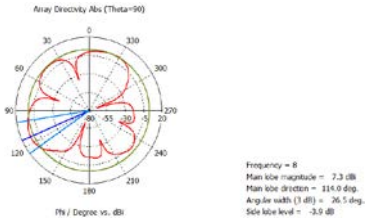


Fig5(o) : Polar diagrams (Theta=90°) at frequency = 8 GHz

3.2 Patch with tree slots

Fig6 illustrates the elementary vertical patch are 36.85 mm long, 29.79 mm wide .And tree slots of 12.21 mm long,9.35 wide . This patch is made of copper. They are fed via holes of 1 mm wide and 2.5mm long. Via holes are connected to the feeding network.

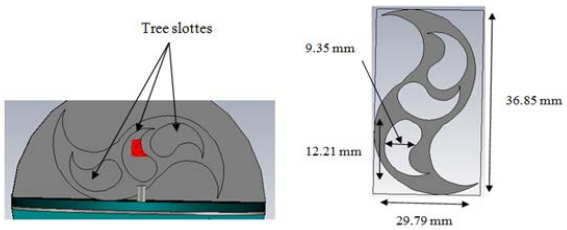


Fig 6: Vertical patch used with tree slots

3.2.1 Radiation properties of antenna

In Fig7, the computed return loss of antenna with tree slots in vertical patches. The simulated antenna by *CST Microwave Studio* software is well adapted at four resonant frequencies of 4.7 GHz, 5.6 GHz, 6GHz, and 7.4GHz. The reflected power reaches the values of -27.75 dB,-24.6 dB,-15.5 dB and -12.5dB at these resonant frequencies respectively.

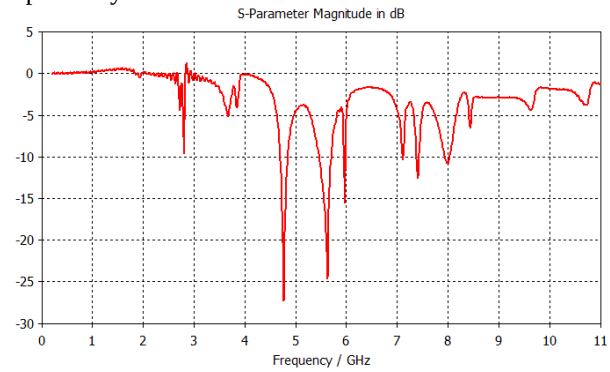


Fig7: Computed return loss of antenna

Fig 8 (a) to Fig 8(o) presents the antenna directivity pattern in two cutting plans at different resonant frequencies.

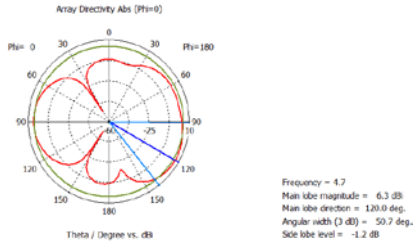


Fig8(a) : Polar diagrams (Phi=0°) at frequency = 4.7 GHz

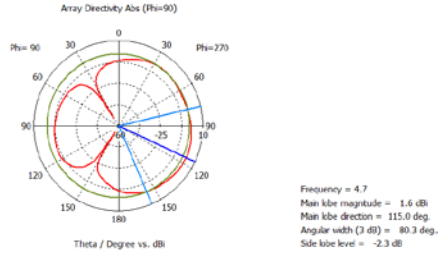


Fig8(b) : Polar diagrams (Phi=90°) at frequency = 4.7 GHz

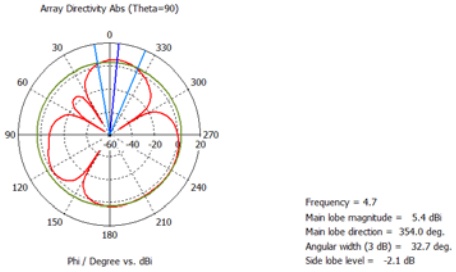


Fig8(c) : Polar diagrams (Theta=90°) at frequency = 4.7 GHz

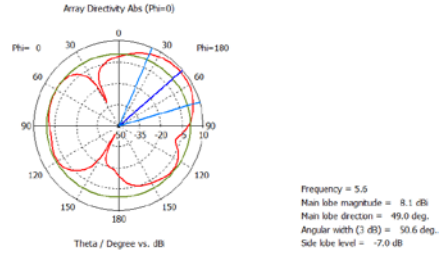


Fig8(d) : Polar diagrams (Phi=0°) at frequency = 5.6 GHz

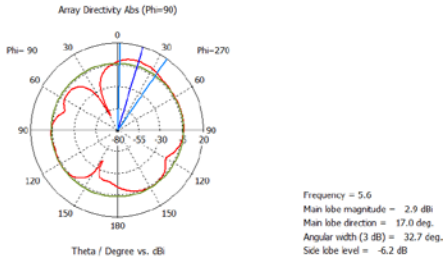


Fig8(e) : Polar diagrams (Phi=90°) at frequency = 5.6 GHz

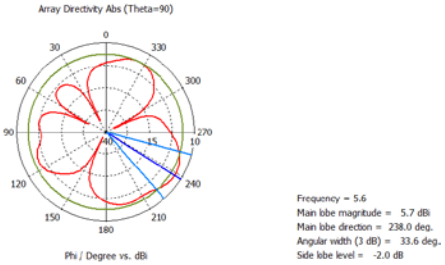


Fig8(f) : Polar diagrams (Theta=90°) at frequency = 5.6 GHz

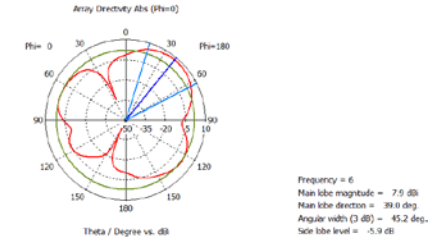


Fig8(g) : Polar diagrams (Phi=0°) at frequency = 6 GHz

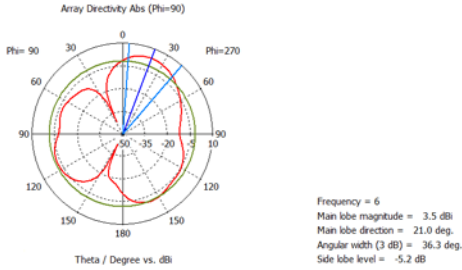


Fig8(h) : Polar diagrams (Phi=90°) at frequency = 6 GHz

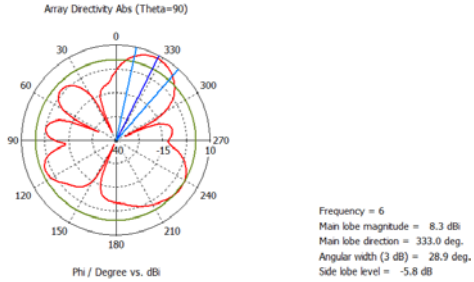


Fig8(i) : Polar diagrams (Theta=90°) at frequency = 6 GHz

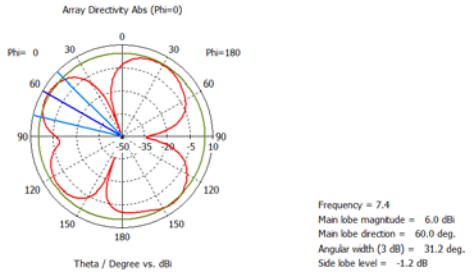


Fig8(j) : Polar diagrams (Phi=0°) at frequency = 7.4 GHz

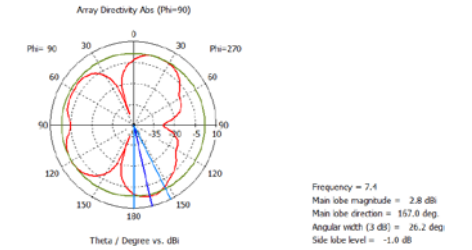


Fig8(k) : Polar diagrams (Phi=90°) at frequency = 7.4 GHz

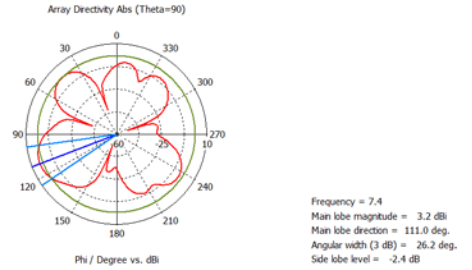


Fig8(l) : Polar diagrams (Theta=90°) at frequency = 7.4 GHz

4. Feeding network

In this section, we propose a microstrip network of the antenna with two 90° hybrid couplers and one 180° hybrid coupler are located in the bottom side of the PCB. As presented the following figure.

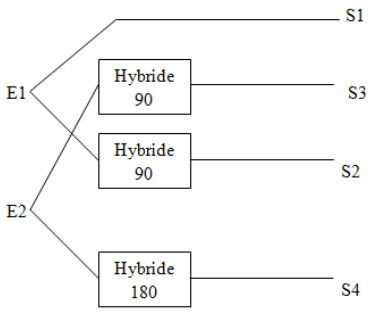


Fig9: Schematic of the feeding network [3]

The circuit layout, as illustrated in Fig 10, was designed using ISIS Proteus. The components are ultra small SMT. The input network is connected through two U-fl coaxial connectors (E1 and E2).

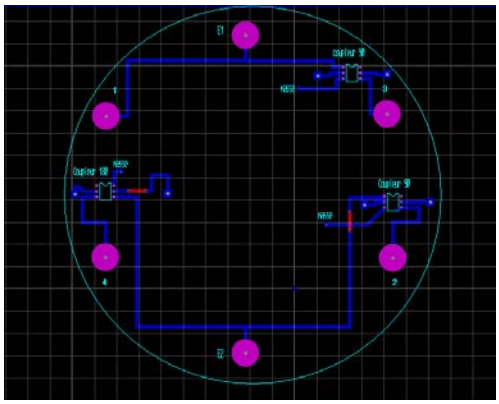


Fig 11: Layout of the microstrip network [4].

The antenna will be positioned in a vertical mode on the PCB.

5. Conclusion

A small multiband miniature antenna with slots on the patches is presented for different application. The simulated results were conducted using the CST Microwave Studio. Furthermore the proposed antenna has a near isotropic pattern for several frequency bands signifying that the proposed antenna is suitable for using in different field in communications. In addition, key advantage of the proposed antenna is simplicity of designing, simple structure, and cost-effective to manufacture.

References

[1] Prapoch Jirasakulporn,'' Multiband CPW-Fed Slot Antenna with L-slot Bowtie Tuning Stub'', World Academy of Science, Engineering and Technology 48 2008.

[2] Mathieu Huchard , Christophe Delaveaud and Smail Tedjini , ''Characterization of the Coverage Uniformity of an Antenna based on its Far-Field'',IEEE 2005 .

[3] Mathieu HUCHARD ,'' Caractérisation et Conception d'Antennes Isotropes Miniatures pour Objets Communicants'', préparée au Laboratoire d'Electronique et de Technologie de l'Information du CEA Grenoble dans le cadre de l'Ecole Doctorale .

[4] Jean-Christophe MICHEL, ''www.gecif.net'' , ISIS Proteus ,July 2011 .

[5] Ross Kyprianou, Bobby Yau , Aris Alexopoulos , Akhilesh Verma and Bevan D. Bates ,'' Investigations into Novel Multi-band Antenna Designs'', Defence Science and Technology Organisation ,PO Box 1500,Edinburgh South Australia 5111 .

[6] Hiroyuki Tamaoka , Hiroki Hamada and Takahiro Ueno,''A Multiband Antenna for Mobile Phones'' , Furukawa Review, No. 26 2004.

[7] H. F. AbuTarboush, H. S. Al-Raweshidy and R. Nilavalan,'' Multi-Band Antenna for Different Wireless Applications'', November 29, 2009 at 15:01 from IEEE Xplore .

[8] Huchard, M., Delaveaud, C.,Tedjini, S,''Miniature Antenna for Circularly Polarized Quasi Isotropic Coverage'', Antennas and Propagation, 2007. IEEE 2007 .

[9] Lev Pazin, Aleksey Dyskin, and Yehuda Leviatan,'' Quasi-Isotropic X-Band Inverted-F Antenna for Active RFID Tags'', IEEE ANTENNAS AND WIRELESS PROPAGATION LETTERS, VOL. 8, 2009 .

[10] Sami HEBIB,''Nouvelle topologie d'antennes multi-bandes pour applications spatiales'', Délivré par l'Université Toulouse III - Paul Sabatier .

[11] Yue Gao, ''Characterisation of Multiple Antennas and Channel for Small Mobile Terminals'', Department of Electronic Engineering Queen Mary, University of London,United Kingdom,June 2007

[12] Huan-Chu Huang , Xiaojing Xu, and Yuanxun Ethan Wang ''Dual-Band Isotropic Radiation Patterns from a Printed Electrically Small Loop-Loaded Dipole Antenna'', Department of Electrical Engineering, University of California at Los Angeles, 405 Hilgard Ave., Los Angeles, CA90095-1594, U.S.

[13] Zhi Ning Chen, Kazuhiro Hirasawa, Kwok-Wa Leung,, and Kwai-Man Luk, , ''A New Inverted F Antenna with a Ring Dielectric Resonator'', IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, VOL.48, NO. 4, JULY 1999 .

Abdellatif BERKAT was born in Algeria in 1987. He obtained his Master's Degree in Telecommunications, from Abou Bekr Belkaid University, Tlemcen, Algeria, in 2010. Abdellatif BERKAT is interested in the following topics: antenna design, algorithmic and programming theories, optimization algorithms, development of artificial intelligence methods. Abdellatif BERKAT is a doctorate student in the same university working on antenna design.

Noureddine Boukli-Hacene was born in 1959 in Tlemcen, Algeria. He received his Diplome d'Etudes Approfondies in microwave engineering (DEA Communications, Optiques et Microondes) and his Doctorate Degree in electrical engineering from Limoges University, France and from the National Center of Spatial Studies (Centre National d'Etudes Spatiales) in Toulouse, France, in 1982 and 1985 respectively. Recently, he was appointed as a lecturer at the University of Tlemcen. His research interests include, among others, microstrip antennas and microwave circuits.

Performance Evaluation and Analytical Validation of Internet Gateway Discovery Approaches in MANET

Rakesh Kumar[1], Anil K. Sarje[2] and Manoj Misra[3]

**[1] Department of Computer Science and Engineering, M.M.M. Engineering College
Gorakhpur, Uttar Pradesh, India-273010**

**[2] Department of Electronics and Computer Engineering, Indian Institute of Technology
Roorkee, Uttarakhand, India-247667**

**[3] Department of Electronics and Computer Engineering, Indian Institute of Technology
Roorkee, Uttarakhand, India-247667**

Abstract

The integration of MANET and Internet extends the network coverage and also increases the application domain of the MANET. The connection of ad hoc networks to the Internet is established via Internet gateways, which acts as a bridge between them. One of the key overhead components affecting the overall performance of this integration is the discovery and selection of Internet gateways as discovery time and handover delay have strong influence on packet delay and throughput. In this paper, the three Internet gateway discovery approaches have been implemented and then the impact of node mobility for two different cases have been examined in terms of performance metrics throughput, end-to-end-delay and routing overhead using network simulator NS2. Our simulation results reveal that the reactive Internet gateway discovery approach scale poorly with increase in number of traffic sources and node mobility to access Internet as compared to the proactive and hybrid gateway discovery approaches. However, reactive gateway discovery results higher throughput and lower end-to-end delay for the same situation than proactive and hybrid approaches. Hybrid Internet gateway discovery approach performance was always observed in between reactive and proactive approaches. The simulation results have also been analytically verified.

Keywords: *Mobile ad hoc network (MANET), Internet gateway discovery, Perormance analysis, AODV, NS2, Internet.*

1. Introduction

MANET applications need a connection to the world wide Internet [1]. For instance members of a conference, which have configured an ad hoc network to exchange information among each other, may need a connection to the Internet to download their emails. For such a scenario, integration of the Internet and the MANET is required. In order to realize such an interworking, an access point, i.e., Internet gateway, is required which has both wired and wireless interfaces. The challenge in interconnecting ad hoc networks to Internet stems from the need to inform ad hoc nodes about available Internet gateways while making a minimal consumption of the scarce network resources. So, an efficient Internet gateway discovery approach for ad hoc networks becomes one of the key elements to enable the use of hybrid ad hoc networks in future mobile and wireless networks. Due to the multi-hop nature of MANET, there might be several reachable Internet gateways for a mobile node at some point of time. If a mobile node receives Internet gateway advertisements from more than one Internet gateway, it has to decide which Internet gateway to use for its connection to the Internet. Several Internet gateway discovery approaches of interconnectivity between mobile ad hoc networks and Internet have been proposed in the literature. However, a comprehensive performance evaluation and comparative analysis of these approaches have not been performed yet. A comprehensive evaluation and performance comparison of Internet gateway discovery approaches in different scenarios will enable one to design and choose a proper Internet gateway discovery approach. This paper sheds some light onto the performance implications of the main features of each approach, presenting simulation results, which provide valuable information to MANET-Internet integration designers. Firstly, we introduce the three existing Internet gateway discovery approaches [2,3,13] and then, based on the simulation results with NS2 [4], we give a detailed comparison and analysis in various network scenarios. In this paper, we investigate the impact of traffic sources and mobility in terms of performance metrics throughput, end-to-end delay, and

routing overhead on the three Internet gateway discovery approaches. We also compare the routing overhead obtained through our simulation with routing overhead computed through analytical model in the same scenario proposed by Ruiz et al. [5] for the three Internet gateway discovery approaches. Figure 1 shows an interworking scenario [1,6,7] in which a mobile node from ad hoc domain wants to communicate with a fixed node on the Internet.

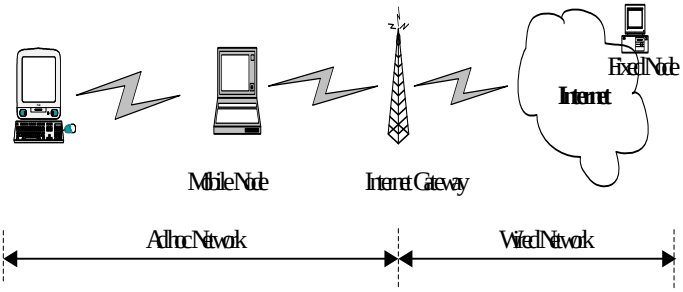


Figure 1: Internet access for ad hoc networks

The remainder of the paper is organized as follows: Related work about Internet gateway discovery approaches and their performance is presented in Section 2. We present the simulation environment, simulation results and its analysis obtained under various conditions, i.e., varying mobility in Section 3. Validation of our simulation results with analytic model has been presented in Section 4. Finally, the paper ends with concluding remarks in Section 5.

2. Related Work

The proposal by Broch et al. [17] is based on integration of MANET with Mobile IP using a source routing protocol. They introduced a border router or gateway, which has two interfaces. Routing on Internet gateway's interface internal to the ad hoc network is accomplished using dynamic source routing (DSR) [18] protocol, while its interface connected to the Internet is configured to use normal IP routing mechanisms. Mobile nodes in an ad hoc network are assigned home addresses from a single network. The nodes within range of the foreign agent act as gateways between the ad hoc network and the Internet. As a reactive approach, foreign agent discovery is only done when required. Traditional IP routing is used on the Internet side, while within MANET, DSR protocol is used. Foreign agents are responsible for connecting the ad hoc network with the Internet.

Hamidian et al. [8] gave a solution, which provides Internet connectivity to ad hoc networks by modifying the AODV routing protocol. An "I" flag is added as an extension to AODV RREQ and RREP to locate the fixed node. If a mobile node fails to receive any corresponding route replies after one network-wide search, it assumes that the destination is a fixed node and is located in the Internet. Thus, it delivers the packets through an Internet gateway. Three methods of gateway discovery for a mobile node to access the Internet are provided: proactive, reactive and hybrid approach. All of them are based on the number of physical hops to gateway as the metric for the gateway selection.

In [11] the scalability of both approaches (proactive and reactive) is compared with respect to the number of Internet gateways by Ghassemian et al. The fixed access network together with the ad hoc fringe constitutes a multihop access network. AODV protocol manages routing in the ad hoc domain. The simulation results show that the proactive approach is more advantageous because the packet delivery ratio is higher and, although the signaling overhead is larger too, it is reduced for a higher number of Internet gateways, because the amount of periodical gateway advertisements is increased but more data packets are transmitted successfully. The hybrid gateway discovery approach is also compared. The hybrid gateway discovery represents a balance between the reactive and the proactive approaches when the number of Internet gateways increases is also reduced.

El-Moshrify et al. [15] proposed a solution in which mobile nodes can access the Internet via a stationary gateway node or access point. Three proposed approaches for gateway discovery are implemented and investigated. Also, the effect of the mobile terminals speed and the number of gateways on the network performance are studied and compared. A mobile node uses no load balancing approach to efficiently discover an Internet gateway in this proposal.

Kumar et al. [19] analyzed Internet connectivity of MANETs via fixed and mobile Internet gateways and pointed out limitations in the existing approaches. It provides a good insight to the research community for further modification and review.

Lakhtaria et al. [16] compared the performance of three gateway discovery protocols. The metrics taken for performance comparison were packet delivery ratio (PDR) and routing overhead.

3. Simulation Model and Performance Evaluation

To assess the performance of the three Internet gateway discovery approaches under the same conditions, we implemented them within the network simulator ns-2.34 [4] using Hamidian [8] approach. The Internet gateway selection function uses the criteria of minimum hops to the Internet gateway, in order to get a fair comparison among the three approaches. The simulations

were conducted on an Intel Core 2 Duo processor at 2.53 GHz, 4 GB of RAM running Fedora Core 5.

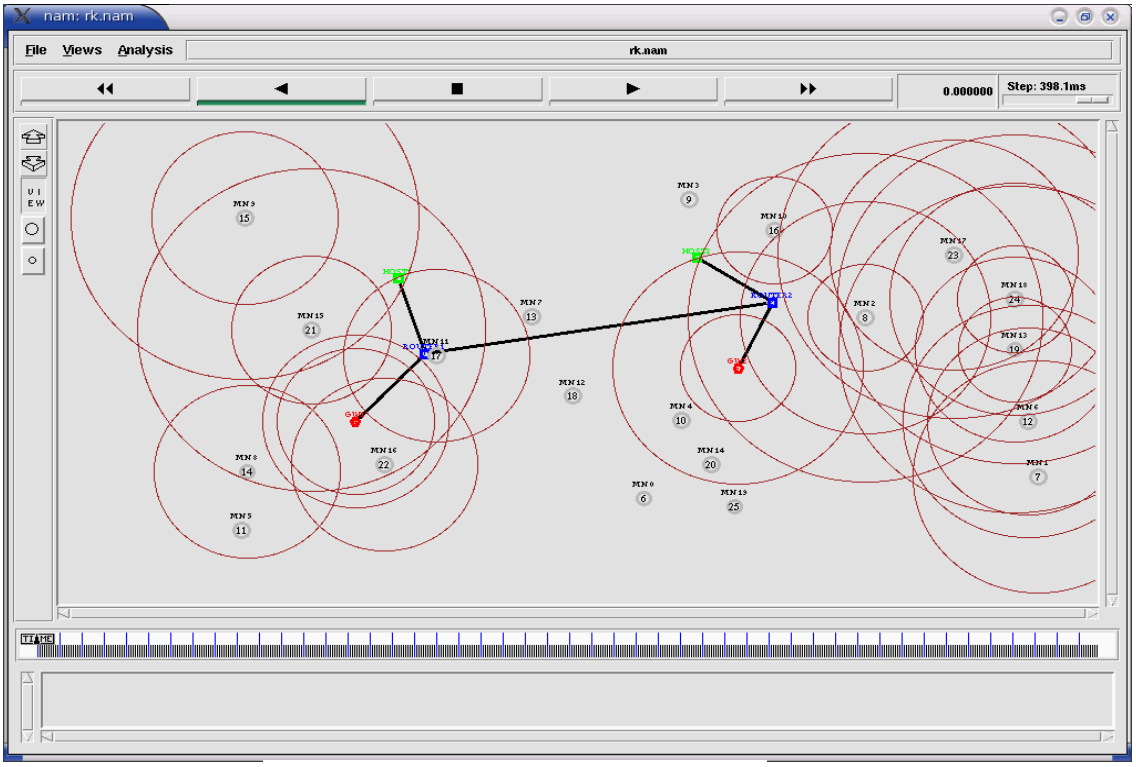


Figure 2: A snapshot of the simulation scenario

3.1 Simulation Model

The studied scenario consists of 20 mobile nodes randomly distributed over an area of 1200×500 m., two fixed hosts host1 and host2 (shown in green color) two routers (shown in blue colors) and two Internet gateways (marked as red colors) as depicted in Figure 2. All fixed links have a bandwidth of 10Mbps, which is enough to accommodate all traffic coming from the mobile nodes. In order to support wireless LAN in the simulator, the Distributed Coordination Function (DCF) of IEEE 802.11 is used as MAC layer protocol. A mobile node uses modified AODV protocol [12] to communicate with its peers and to access wired networks through an Internet gateway. All simulations were run for 500 seconds of simulation time. Two different cases (Case I and Case II) have been considered. In the first case, i.e., Case I, we take three CBR data sources as given in Table I. Mobile nodes MN7, MN12 and MN16 start sending data at t_{SIM} =5 second to host1 through one of the two Internet gateways. We then vary the node mobility as per data given in Table I. The traffic sources connected to mobile nodes MN12 and MN16 keep on sending data at constant rate, i.e., 320 Kbps (packet inter arrival time=0.0125 second, so data rate = (1/. 0125)*512*8=320 Kbps). In this way three different flows (fid=0, fid=1 and fid=2) are active in the network.

3.2 Movement Model

The mobility model used in this study is the Random Waypoint Model [9]. As per this model, a mobile node remains stationary for a specified pause time, after which it begins to move with a randomly chosen speed (0 to 20 m/s) towards a randomly chosen destination within the defined topology. The mobile node repeats the same procedure until the simulation ends. The random speed is chosen to be a value, which is uniformly distributed between a defined minimum and maximum value as given in Table 1. We generated mobile nodes movement pattern by using CMU's movement generator. The command used is:

$./setdest [-n num_of_nodes] [-p pausetime] [-s maxspeed] [-t simtime] \ [-x maxx] [-y maxy] > [outdir/movement-file]

3.3 Communication Model

The communication model is determined by four factors: number of sources, packet size, packet rate and the communication type. We used the CBR (constant bit rate) communication type, which uses UDP (User Datagram Protocol) as its transport protocol. CBR traffic has been used instead of TCP. The reason is that TCP performs poorly in ad hoc network because packets that are lost due to link failure and route changes trigger TCP's congestion avoidance mechanism [10]. Three and six sources are used to generate network traffic (CBR) with sending rate as given in Table I. The packet size of 512 bytes is used throughout the simulation. The traffic connection pattern is generated through CMU's traffic generator (cbrgen.tcl). The main parameters in cbrgen.tcl are "connections" (number of sources) and "rate" (packet rate). So, the command used is:

$ns cbrgen.tcl [-type cbr|tcp] [-nn nodes] [-seed seed] [-mc connections][-rate rate]

Table I: Simulation Parameters for Simulation Model

Parameters	**Value**
Number of mobile nodes	20
Number of sources	3 and 6
Number of gateways	2
Number of fixed nodes	2
Topology size	1200 meters × 500meters
Transmission range	250 meter
Traffic type	Constant Bit Rate (CBR)
Packet sending rate (Kbps) of mobile node MN7	8
Packet sending rate (Kbps) of mobile nodes MN8, MN10, MN12, MN16 and MN20 to host1 or host2	320 (fixed)
Packet Size	512 bytes
Mobile node speed	1,5,10,15 and 20 m/sec
Mobility model	Random Waypoint
Pause time	5 seconds
Link level layer	802.11 DCF
Carrier sensing range	500 meters
Simulation time	500 seconds
Wireless channel bandwidth	2 Mbps
Interface queue limit (wireless node)	50 packets
Interface queue limit (wired node)	50 packets
ADVERTISEMENT_ INTERVAL	5 seconds
ADVERTISEMENT_ZONE	4 hops
Wired link bandwidth	10 Mbps
Buffer management of wired nodes	Drop Tail

3.4 Performance Metrics

In order to investigate the effect of traffic load and mobility on three different gateway discovery approaches, we used the following performance metrics:

Throughput: It is defined as the ratio of total number of data bits (i.e. packets) successfully received at the destination to the simulation time.

End-to-End Delay: It is defined as the delay for sending packets from source node to the fixed host. This metric includes all possible delays caused by buffering during the Internet gateway discovery latency, route discovery latency, queuing at the interface queue, retransmission delays at the MAC layer, and propagation and transfer times.

Routing Overhead: It is defined as the ratio of the AODV packets to the data packets sent and received by all the mobile nodes.

3.5 Simulation Parameters

The common parameters for all the simulations are given in Table I similar with [11].

3.6 Simulation Results And Analysis

We present in this subsection the performance of three Internet gateway discovery approaches for the various metrics presented above.

Effect of Node Mobility

In this sub section, we examine the effect of node mobility on performance metrics throughput and end-to-end delay for the two cases studied earlier. For both the cases, i.e. for Case I and Case II, MN7 sends only at 8 Kbps.

Figure 3 shows the average throughput for CBR traffic at host1 (i.e. node MN7 $\rightarrow$ host1 with flow id 0) for the three Internet gateway discovery approaches for Case I where mobile node speed varies from 1 m/s up to 20 m/s with 5 seconds of pause time. At low speed, the throughputs of the three algorithms are almost similar

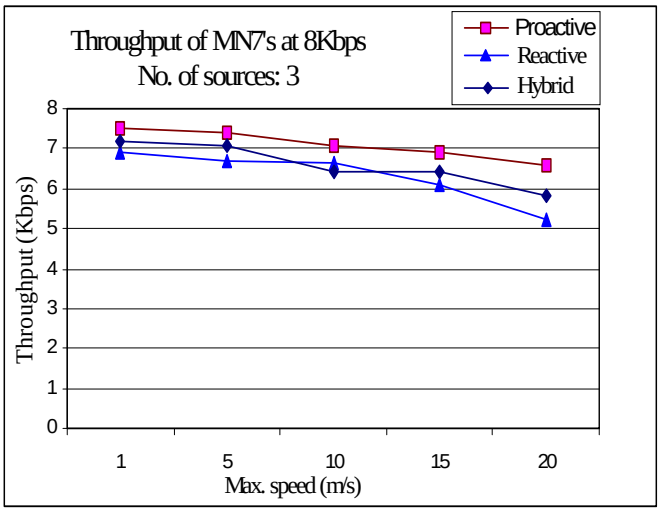


Figure 3: Throughput vs node mobility for Case I (Source: node MN7, Destination: host1, sources: 3)

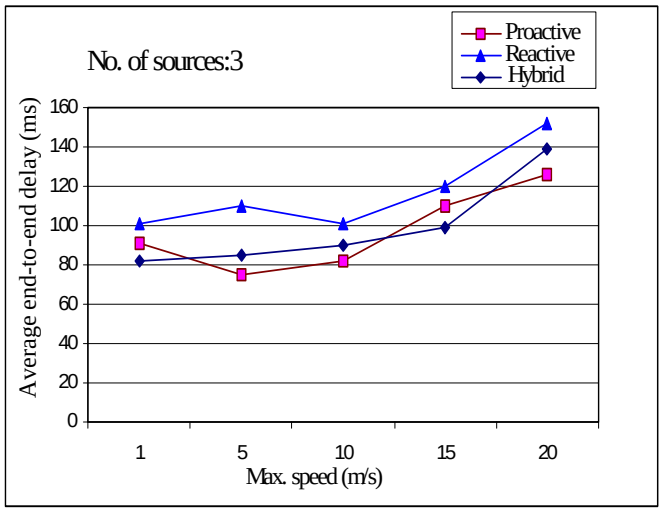


Figure 5: Average end-to-end delay vs node mobility for Case I (Source node MN7, Destination node host1, sources: 3)

and quite good but as node speed increases, due to frequent link changes and connection failures, packet drops occur and throughput starts decreasing. However, the proactive and hybrid approaches have larger throughput than the reactive approach at higher node speed. Reactive discovery results in lower throughput as the source continues to send data packets, which get lost due to link breaks until a route error packet is received by the sending mobile node.

For Case II, throughput of mobile node MN7 decreases with speed. In this case proactive discovery gives lower throughput as compared to others at higher node speeds (Figure 4).

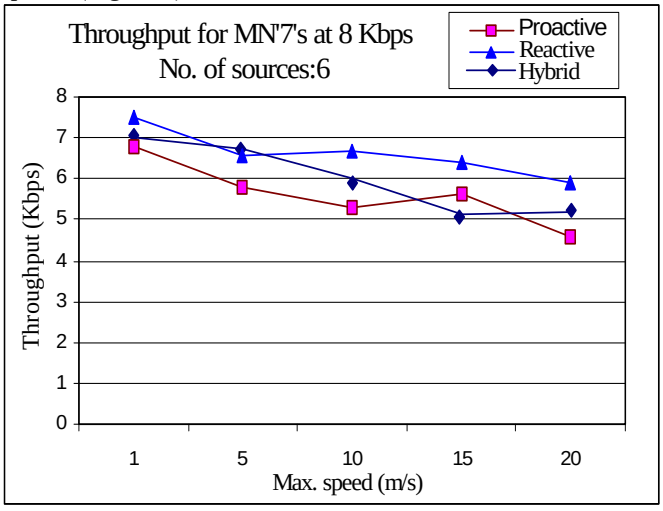


Figure 4: Throughput vs node mobility for Case II (Source: node MN7, Destination: host1, sources: 6)

Performance of hybrid gateway discovery remains in between proactive and reactive for both the cases (see Figure 3 and Figure 4). Moreover throughputs of MN7 are lower in the three gateway discoveries when the traffic sources are increased form 3 to 6. Figure 5 and Figure 6 show average end-to-end delay of MN7 for the three gateway discovery approaches for mobile node speed from 1 m/s up to 20 m/s, pause time of 5 seconds and number of sources 3 and 6. Figure 5 represents the average end-to-end delay for CBR traffic (for flow id 0 between mobile node MN7 and host1) for Case I on the three discovery approaches as mobility increases.

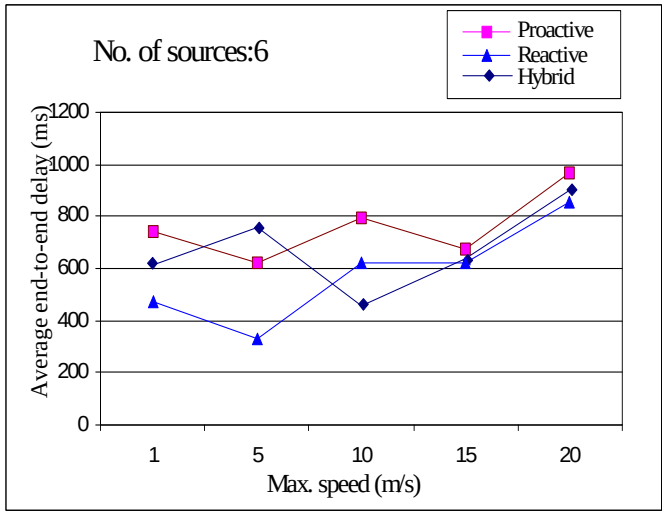


Figure 6: Average end-to-end delay vs node mobility for Case II (Source: node MN7, Destination: host1, sources: 6)

As the figure shows, the average end-to-end delay is lower for the hybrid and proactive approaches than for the reactive approach and increases rapidly with mobile node speed. This is because mobile nodes update their route entries for the gateways more frequently in case of either proactive or hybrid discovery approach which results in shorter and fresher routes. This increases average end-to-end delay in case of reactive discovery approach.

Figure 6 depicts the variation in the average end-to-end delay of packets with mobile node speed for six sources (Case II). In this case, proactive discovery results in higher average end-to-end-delay due to high node mobility compared to reactive and hybrid. But at higher mobility, the difference in end-to-end delay for the three gateway discovery approaches becomes lesser.

4. Validation of Simulation Results with Analytical Model

Ruiz et al. [5] presented an analytic model for the above three Internet gateway discovery approaches for analyzing scalability issue. Gateway discovery overhead is used as performance metric to measure the scalability of an Internet gateway discovery approach. It is the total number of control messages associated with the discovery of an Internet gateway.

This metric gives information about the control overhead to provide Internet connectivity. Table II shows a summary of basic parameters used in the model. Metric chosen for a route to the Internet gateway is the hop count

Table II: Notations used in the derivation

Notations	Meaning
N	Total number of nodes in a square lattice covering a certain area.
N_G	Number of Internet gateways
$N - N_G$	Number of ad hoc nodes
S	Number of active sources communicating with fixed nodes.
t	The time interval during which all sources send CBR traffic to the fixed nodes through Internet gateways
λ_{adv}	The rate at which GWADV messages are being sent out by Internet gateways
λ_{dur}	A parameter used to compute route duration time λ_{dur}

as this metric enables a mobile node to select the nearest Internet gateway to communicate with hosts in the Internet.

4.1 Reactive Gateway Discovery Overhead

In reactive gateway discovery, a source node discovers an Internet gateway reactively. Therefore, in this case gateway discovery overhead includes Internet gateway route request broadcast messages, plus Internet gateway reply messages from every Internet gateway to the source. The overhead of the reactive Internet gateway discovery for one source is given by the following equation [5]

$$R_{overhead} = [F_{overhead} + (R1_{overhead} \times \lambda_{dur} \times t)] \times S \quad (1)$$

where, $F_{overhead}$ gives the number of messages needed to realize that a destination is a fixed node and is given by the following equation

$$F_{overhead} \sum_{j \in (1,3,5,7,30)} N_r(j) \quad (2)$$

Given a broadcast message with time to live (TTL) equal to x, $N_r(x)$ is the number of nodes forwarding this message .

$R1_{overhead}$ represents the overhead of the reactive discovery of the gateway for one source and can be computed by the following equation [5]

$$R1_{overhead} = N + N_G \times \sqrt{N} \quad (3)$$

4.2 Proactive Gateway Discovery Overhead

In proactive approach, Internet gateways periodically broadcast Internet gateway messages (GWADV) to an entire ad hoc network. Therefore, total overhead in number of messages required in this approach can be computed by the following equation

$$P_{overhead} = S \times F_{overhead} + \lambda_{adv} \times t \times (N+1) \times N_G \quad (4)$$

4.3 Hybrid Gateway Discovery Overhead

The hybrid gateway discovery approach has the combined overhead of proactive and reactive approaches. The number of nodes within a scope of s hops from any Internet gateway G_i is given by the following equation

$$N_r^{G_i}(s) = \frac{s \times (s+3)}{2} \quad (5)$$

with $s \in [0, \sqrt{N} - 1]$

The probability of a given ad hoc node receiving a GWADV message from any of the Internet gateways is given by

$$P_c(s) = \frac{\sum_{i=1}^{N_G} N_r^{G_i}(s)}{N - N_G} \quad (6)$$

The overall overhead of the hybrid gateway discovery approach is due to the following overhead:

- overhead to realize that the destinations are outside MANET.
- overhead in broadcasting of GWADV messages over s hops by each Internet gateway.
- overhead needed by those sources not covered by the GWADV messages. These nodes find Internet gateways and create a default route.

Therefore, the total overhead in number of messages required by the hybrid approach can be computed by the following equation

$$H_{overhead} = S \times F_{overhead} + \lambda_{adv} \times t \times (N_r^G(s) + 1) \times N_G + R1_{overhead} \times \lambda_{dur} \times t \times S \times (1 - P_c(s)) \quad (7)$$

The Internet gateway discovery overheads of the three Internet gateway discovery approaches when the number of active traffic sources are 3 and 6 are computed using the approaches obtained through analytical model and simulation above analytic model. The results obtained with analytic model and simulation for the scenario considered (with parameters taken from Table I, and t=500 s, λ_{adv} =1/5 as Internet gateway advertisement interval is 5 sec) is listed in Table III. These simulated results are compared with the analytical results in Table III. We can see that all the figures are quite similar, taking into account that the model and the simulated environment have many differences (simulated area, mobility, MAC layer, etc), so some deviation is expected.

From the analytical results obtained from analytical model given by Ruiz et al. [5]) and with our simulation results, it can be concluded that as the number of traffic sources and mobility increase, reactive approach incurs higher overhead than proactive and hybrid approaches. Therefore the reactive approach shows poor scalability as number of sources connecting to the Internet increases. Hybrid gateway discovery approach incurs minimum overhead. This validates our simulation results for scalability issue.

5. Conclusions

In this paper, we considered Internet connectivity of ad hoc networks via Internet gateways. AODV routing protocol for ad hoc networks has been modified to offer enhanced Internet connectivity and then we investigated in depth the effect of traffic sources and node mobility on the three Internet gateway discovery, viz. reactive, proactive and hybrid for providing inter-connectivity between ad hoc networks and Internet. The performance metrics chosen are throughput, average end-to-end delay and routing overhead. To assess the performance of this idea, simulation has been carried out using NS2 Simulator [4] for two different cases (number of active sessions, i.e., for 3 and 6 sources). At low mobility, the performance of proactive and hybrid gateway discovery is better as compared to reactive discovery. They result in higher throughput, lower end-to-end delay compared to reactive approach. But as number of sources and node mobility increases, the reactive gateway discovery outperforms proactive and gives similar performance with hybrid discovery approach. Reactive gateway gives higher throughput and lower end-to end delay than proactive approach. However reactive approach shows poor scalability as number of sources connecting to Internet increases which is confirmed by comparing routing overheads obtained through our simulation and routing overheads computed through analytical model [5]. Performance of hybrid gateway discovery approach always remains in between reactive and proactive approaches. However, the overall performance of the three Internet gateway discovery approaches are very much dependent on the prevailing network conditions.

Table III: Routing overheads of the three Internet gateway discovery

No. of sources	Overheads of Internet gateway discovery approaches (Number of messages)					
	Proactive approach		Reactive approach		Hybrid approach	
	Simulation	Analytical	Simulation	Analytical	Simulation	Analytical
3	12605	14475	13989	14668	10078	8205
6	18283	20670	25934	29336	20384	11012

References

[1] J. Farooq, "Mobility and Internet Connectivity in Mobile Ad hoc Networks," Proceedings Umea's 9th Students Conference in Computing Science, pp. 269-283, June 3-4, 2005.

[2] J. Xi and C. Bettstetter, "Wireless Multi-hop Internet Access: Gateway Discovery, Routing, and Addressing," Proceedings of International Conference on Third Generation Wireless and Beyond (3Gwireless), San Francisco, USA, May 2002.

[3] R. Wakikawa, J. Malinen, C. Perkins, A. Nilsson and A. Tuominen, "Internet Connectivity for Mobile Ad hoc Networks," Internet-Draft, draft-wakikawa- manet-global6-02.txt, November 2002, Work in progress.

[4] The Network Simulator NS-2. http://www.isi.edu/nsnam/ns

[5] P. M. Ruiz and A. F. Gomez-Skarmeta, "Maximal Source Coverage Adaptive Gateway Discovery for Hybrid Ad hoc Networks," Lecture Notes in Computer Science (LNCS) 3158, pp. 28-41, 2004.

[6] A. Shaikh, J. Rexford and K. G. Shin, "Load-Sensitive Routing of Long Lived IP Flows," Proceedings of the ACM SIGCOMM Conference on Communication Architectures, Protocols and Applications, Cambridge, MA, pp. 215-226, 1999.

[7] R. Wakikawa, J. T. Malinen, C. E. Perkins and A. Nilsson, "Global Connectivity for IPv6 Mobile Ad hoc Networks" In IETF Internet Draft 2003. http://www.wakikawa.net/Research/paper/draft/manet/draft-wakikawa-manet-globalv6-03.txt

[8] A. Hamidian, U. Korner and A. Nilsson, "A Study of Internet Connectivity for Mobile Ad hoc

Networks in NS2", Department of Communication Systems, Lund Institute of Technology, Lund University, January 2003.

[9] E. Hyytia and H. Koskinen, "Random Waypoint Model in Wireless Networks," Networks and Algorithms: Complexity in Physics and Computer Science, Helsinki, pp 16-19, 2005.

[10] J. Schiller, "Mobile Communication," 2nd Edition, Addison Wesley, 2003.

[11] M. Ghassemian, P. Hofmann, C. Prehofer, V. Friderikos and H. Aghvami, "Performance Analysis of Internet Gateway Discovery Protocols in Ad hoc Networks," IEEE Wireless Communications and Networking Conference (WCNC), Atlanta, Georgia, USA, 2004.

[12] C. E. Perkins, E. M. Belding-Royer and S. Das, "Ad hoc On-Demand Distance Vector (AODV) Routing," draft-perkins-manet-aodvbis-00.txt, Internet Draft, 19 October 2003.

[13] Rakesh Kumar, Manoj Misra, and Anil K. Sarje, " A Simulation Analysis of Gateway Discovery for Internet Access in Mobile Ad hoc Networks," International Journal of Information Processing, Vol. 2, Issue 2, pp. 52-64, 2008.

[14] P. Ruiz, A. Gomez-Skarmeta, " Enhanced Internet Connectivity for Hybrid Ad hoc Networks through Adaptive Gateway Discovery," Proceeding of the 29th Annual IEEE International Conference on Local Computer Networks, LCN 04, Tampa, Florida, November 2004.

[15] H. El-Moshrify, M. A. Mangoud and M. Rizk, "Gateway Discovery in Ad hoc On-Demand Distance Vector (AODV) Routing for Internet Connectivity," 24th National Radio Science Conference (NRSC 2007), Faculty of Engineering, Alexandria University Alexandria 21544, Egypt, March 13-15, 2007.

[16] K. I. Lakhtaria and B. N. Patel, "Comparing Different Gateway Discovery Mechanism for Connectivity of Internet & MANET," International Journal of Wireless Communication and Simulation, Vol. 2, No. 1, pp. 51–63, 2010.

[17] J. Broch, D. A. Maltz and D.B. Johnson, "Supporting Hierarchy and Heterogeneous Interfaces in

Multi-Hop Wireless Ad hoc Networks," Proceedings of the IEEE International Symposium on Parallel Architectures, Algorithms, and Networks, Perth, Western Australia, pp. 370-375, June 1999.

[18] D. B. Johnson, D. A. Maltz and J. G. Jetcheva, "The Dynamic Source Routing Protocol for Mobile Ad hoc Networks," IETF Internet Draft, draft-ietf-manet-dsr-07.txt, work in progress, 2002.

[19] Rakesh Kumar, Anil K. Sarje, Manoj Misra, "Review Strategies and Analysis of Mobile Ad Hoc Network-Internet Integration Solutions," IJCSI International Journal of Computer Science Issues, Volume 7, Issue 4, July 2010.

Dr. Rakesh Kumar received PhD in Computer Science & Engineering from IIT Roorkee, India in 2011, M.E. in Computer Engineering from S.G.S. Institute of Technology and Science Indore, India in 1994 and B. E. in Computer Engineering (First class with honours) from M. M. M. Engineering College, Gorakhpur, UP, India in 1990. Dr. Kumar is in teaching, research & development since 1992 and is presently working as an Associate Professor in the Department of Computer Science and Engineering, M.M.M. Engineering College Gorakhpur-India. He has published many research papers in many refereed International/National Journals and International Conferences. He had been awarded for Best Research paper in an International conference. He is Fellow of IETE and IE and also member of CSI and ISTE. His research interests are in Mobile & Distributed Computing, Mobile Ad hoc Routing, Quality of Service Provisioning, MANET-Internet Integration, Sensor Networks and Performance Evaluation.

Dr. Anil K. Sarje is Professor in the department of Electronics & Computer Engineering at Indian Institute of Technology Roorkee, India. He received his B.E., M.E. and PhD degrees from Indian Institute of Science, Bangalore in 1970, 1972 and 1976 respectively. He served as Lecturer at Birla Institute of Technology & Science, Pilani, for a short period before joining University of Roorkee (now Indian Institute of Technology Roorkee) in 1987. Prof. Sarje has supervised a large number of M.Tech. Dissertations and guided several Ph.D. theses. He has published a large number of research papers in the International and National journals and conferences. He has also served as referee for many reputed Journals like IEE Proceedings, IEEE Transaction on Reliability, Electronics Letters, etc. He has been on a number of AICTE and DOEACC committees. He was a member of All India Board of Information Technology during years 2000-2003. He is a senior member of the Institute of Electrical and Electronics Engineers (IEEE). His research interests include Distributed Systems, Computer Networks, Real Time Systems and Network Security.

Dr. Manoj Misra is a Professor in the department of Electronics & Computer Engineering at Indian Institute of Technology Roorkee, India. He received his B.Tech. degree in 1983 from H.B.T.I., Kanpur and M.Tech. from University of Roorkee in 1986. He did his Ph.D. from Newcastle upon Tyne, UK and joined Electronics & Computer Engineering Department, University of Roorkee (now Indian Institute of Technology Roorkee) in August 1998 as Assistant Professor. Before joining University of Roorkee, he worked in DCM, CMC Ltd., New Delhi, H.A.L. Kanpur and H.B.T.I. Kanpur. He has completed an AICTE funded project "A CORBA framework for distributed mobile applications", as a co- Investigator with Prof. R. C. Joshi. Prof. Misra has supervised a large number of M. Tech. Dissertations and guided several Ph.D. Theses. He has published a large number of research papers in International and National journals and conferences. He is a member of the Institute of Electrical and Electronics Engineers (IEEE). His research interests include Distributed Computing and Performance Evaluation.

Scalable Symmetric Key Cryptography Using Asynchronous Data Exchange in Enterprise Grid

Medhat Awadallah[1] and Ahmed Youssef[2]

**[1] Electrical and Computer Engineering Dept, Sultan Qaboos University
Muscat, Oman**

**[2] Information Systems Department, King Saud University
Riyadh, 11543, KSA**

Abstract

Symmetric key cryptography is one of the most critical computing problems that need high performance computing power resources. The use of large key sizes and complex encryption/decryption algorithms to achieve unbreakable state has led to an increased time computational complexity. Traditionally, this problem is solved in the grid environment by partitioning data streams into several blocks of a predefined size. This is done while sequentially reading the data from the raw data file. The grid manager node then takes the responsibility of passing these blocks to the executer nodes where different blocks are processed separately and simultaneously. Although this technique allows parallel processing to speed up the encryption/decryption process, creating blocks by sequentially reading the data file and distributing these blocks on executers synchronously by the central manager node is a poor technique and a source of delay. In this paper, we present a novel approach that tackles this problem by allowing executers to access data file at random and asynchronously exchange the blocks among them, thereby, delay is significantly reduced and data size can be scaled up. In order to show the merit of our approach experiments have been conducted through a system-level middleware for grid computing called Alchemi. The results show a remarkable performance enhancement in our approach over traditional approaches in terms of speed.

Keywords: Grid computing, Grid Middleware, Alchemi, Data Encryption/Decryption, Symmetric Key Cryptography.

1. Introduction

The concept of grid computing is gaining popularity with the emergence of the Internet as a medium for global communication and the wide spread availability of powerful computers and networks as low-cost commodity components [1]. The computing resources and special class of scientific devices or instruments are located across various organizations around the globe. These resources could be computational systems (such as traditional supercomputers, clusters [2], or even powerful desktop machines), special class of devices (such as sensors, radio telescope, and satellite receivers), visualization platforms, or storage devices. A number of applications need more computing power than can be offered by a single resource/reasonable time and cost. This promoted the exploration of logically coupling geographically distributed high-end computational resources and using them for solving large-scale problems. Such emerging infrastructure is called computational (power) grid [3]. Computational grids are expected to offer dependable, consistent, pervasive, and inexpensive access to high-end resources irrespective of their physical location and the location of access points [3].

The grid must be designed and created in such a way that their components (fabric, middleware, and higher-level tools) and applications handle the key design issues in a coordinated manner. For instance, grid middleware offers services for handling heterogeneity, security, information, allocation, and so on. Higher level tools, such as resource brokers, support dynamic adaptability through automatic resource discovery, trading for economy of resources, resource acquisition, scheduling, the staging of data and programs, initiating computations, and adapting to changes in the grid status [4]. In addition, they also need to make sure that domain autonomy is honored but still meets user requirements such as quality of service in coordination with other components.

Symmetric key cryptography is one of those complex large-scale problems that need high computing power to be solved efficiently. Cryptanalysis on this problem is

encouraging the use of larger key sizes and complex algorithms to achieve an unbreakable state [5]. However, this leads to an increase in computational complexity. Therefore, many researchers investigated the deployment of high performance computing approaches such as grid computing, cluster computing and Peer-to-Peer (P2P) to develop efficient and cost-effective symmetric key cryptography schemes. By utilizing these approaches, the performance of symmetric key cryptography can be improved through parallel execution [5].

Traditionally, this problem is solved in the Grid environment by partitioning data streams into several blocks of a predefined size [5, 14]. This is done while sequentially reading the data from the raw data file. The grid manager node then takes the responsibility of assigning these blocks to the executer nodes where different blocks are processed separately and simultaneously. Although this technique allows parallel processing to speed up the encryption/decryption process, creating blocks by sequentially reading the data file and distributing these blocks on executers synchronously by the central manager node is poor technique and a source of delay. In this paper, we present a novel approach that tackles this problem by allowing executers to access data file at random and asynchronously exchange data blocks among them. The proposed approach is faster and more scalable than traditional approaches since it avoids the delay occurs due to partitioning the data into blocks by the grid application while reading the file and passing large sets of data by the manager to the executers. The validity and the feasibility of the proposed approach is examined through a system level middleware for creating grid computing environment called Alchemi. Experiments show a remarkable performance enhancement in our approach over traditional approaches.

The rest of this paper is organized as follows: in section 2 we outline background information. Section 3 presents the open source, Alchemi, which provides the middleware for creating an enterprise grid-computing environment. Section 4 presents DES (Data Encryption Standard); Encryption and Decryption using Alchemi grid computing framework. Our proposed approach is presented in Section 5. Section 6 presents performance evaluation experiments conducted through Alchemi and discusses the results. Finally, section 7 gives our conclusions.

2. Background

In order to meet the increasing demand of large-scale scientific computation in the fields of life sciences, biology, physics, and astronomy, the notion of "computational grid" was proposed in mid 1990s [6]. It has been observed that computers (such as PCs, work-stations, and clusters) in the Internet are often idle. Grid computing aims to integrate idle computational power over the Internet and provide powerful computation capability for users all over the world [7]. Since a grid connects numerous geographical distributed computers fashion, an important issue is how to evenly distribute submitted tasks to nodes. This is a load balancing problem, one of the scheduling problems on the grid. By solving this problem, the computational resources of the grid can optimally be utilized. To perform grid computation, the process must be divisible into several sub-processes and run in parallel. The following are some of famous projects that have been designed for grid computation.

The human genome is composed of 24 distinct chromosomes with about 3 billion DNA base pairs organized into 20,000~25,000 genes [8]. To identify these genes and determine the sequences of 3 billion DNA base pairs, running a computer simulation would be expensive and time consuming. Based on computational grid, the Human Genome Project was completed in 2003, three years ahead of the target goal. After the Human Genome Project was completed, scientists wanted to understand the function of human proteins, which affect human health, to discover the cure for diseases such as AIDS and cancer. Human Proteome Folding (HPF) project was started and ran on two computational grids [9].

Chemical reactions or molecular behavior can be huge and complicated processes. Some chemistry problems, like quantum mechanics, would take hundreds of years to simulate on a personal computer. Computational Chemistry Grid [10] is one of the most important virtual organizations, which provides all necessary software and resources for computational chemistry. Searching for extraterrestrial intelligence (SETI), is a compelling scientific research that utilizes grid computation technology to analyze space-based radio signals collected from a radio telescope, at Arecibo, Puerto Rico [12].

Grid computation is not only used in science, but also in business computation, where all corporate resources can be pooled so they can be processed efficiently in parallel, according to the business demand. The Oracle 10g [11] runs all database systems in a virtual environment (grid) where all systems are considered a resource pool, using resources efficiently and dynamically for business needs. Grid computation can also be used in financial modeling,

earthquake simulation, and climate/weather modeling, which are complex processes requiring an intricate infrastructure. A dynamic grid environment, which can perform parallel processing under a collaborative network, must be created to deliver the information. A number of projects worldwide are actively exploring the development of grid computing technology. They include Globus [13], Legion [15], NASA Information power grid [16], and Condor [17].

3. Windows-based grid computing framework (Alchemi)

The Alchemi grid-computing framework was conceived with the aim of making grid construction and development of grid software as easy as possible without sacrificing flexibility, scalability, reliability and extensibility. The key features supported by Alchemi are [18,19]:

- Windows based machine with .NET grid computing framework;
- Internet-based clustering of desktop computers without a shared file system
- Federation of clusters to create hierarchical, cooperative grids
- Dedicated or non-dedicated (voluntary) execution by clusters and individual nodes
- Object-oriented grid thread programming model (fine-grained abstraction)
- Web services interface supporting a grid job model (coarse-grained abstraction) for cross-platform interoperability (e.g., for creating a global and cross-platform grid environment via a custom resource broker component).

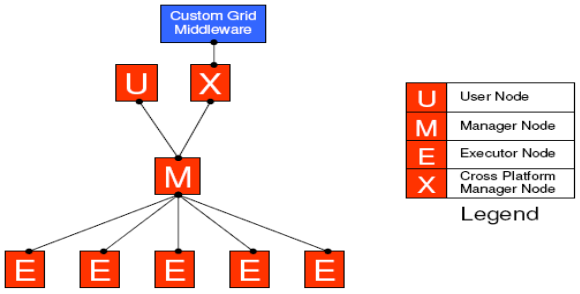
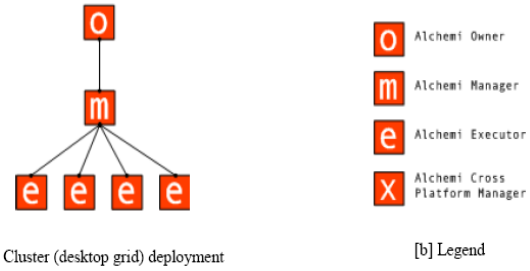


Fig. 1 Distributed components and their relationships [19]

Alchemi's distributed components consist of four types of nodes (or hosts) that take part in enterprise grid construction and application execution. These nodes include: User node, Manager node, Executer node and Cross platform Manger node, Fig. 1. An Alchemi enterprise grid is constructed by deploying a Manager node and one or more Executor nodes configured to connect to the Manager. One or more Users can execute their applications by connecting to the Manager. An optional component, the Cross Platform Manager, provides a web service interface to custom grid middleware. These components allow Alchemi to be utilized to create different grid configurations, which are desktop cluster grid, multi-cluster grid, and cross-platform grid (global grid). According to [19], these are described as follows:

Cluster (Desktop Grid): is the basic deployment scenario, a cluster (as shown in Fig. 2) consists of a single Manager and multiple Executors that are configured to connect to the Manager. One or more Owners can execute their applications on the cluster by connecting to the Manager. Such an environment is appropriate for deployment on Local Area Networks as well as the Internet.

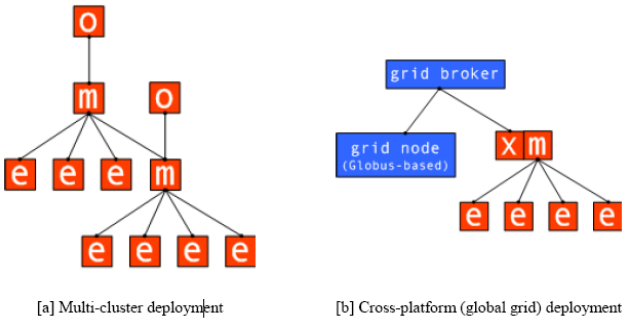


[a] Cluster (desktop grid) deployment [b] Legend

Fig. 2 Cluster (desktop grid) deployment [19]

Multi-cluster environment: is created by connecting Managers in a hierarchical fashion, Fig. 3.a. As in a single-cluster environment, any number of Executors and Owners can connect to a Manager at any level in the hierarchy. An Executor and Owner in a multi-cluster environment connect to a Manager in the same fashion as in a cluster and correspondingly their operation is no different from that in a cluster.

Global Grid: the cross platform manager is used to construct a grid conforming to the classical global grid model, Fig. 3.b. A grid middleware component such as a broker can use the Cross-Platform Manager web service to execute cross-platform applications (jobs within tasks) on an Alchemi node (cluster or multi-cluster) as well as resources grid-enabled using other technologies such as Globus.



[a] Multi-cluster deployment [b] Cross-platform (global grid) deployment

Fig.3: Alchemi deployment in [a] multi-cluster [b] global grid environments [19]

4. DES Encryption/Decryption using Alchemi grid computing framework

In this paper, we are concerned with symmetric key encryption algorithms such as DES and RC4 [5] as a grid application that runs under Alchemi framework. These algorithms are extremely fast (compared to public-key algorithms) and are well suited for performing cryptographic transformations on large streams of data. Typically, these algorithms are used to encrypt one block of data at a time. Block ciphers cryptographically transform an input block of n bytes into an output block of encrypted bytes. The Enterprise Grid Middleware (Alchemi) is used to solve the symmetric key cryptography problem as shown in Fig. 4. Alchemi provides a Software Development Kit (SDK) that can be used by developers to develop grid applications. The SDK includes a Dynamic Link Library (DLL) that supports object oriented programming model for multithreaded applications.

A grid application, called GridCryptoGraphy, has been built on top of Alchemi middleware grid environment as shown in Fig. 5. In this application, three main classes have been developed. The first class (GridCryptForm) is the interface to control and monitor the progress of the encryption and decryption process. It is also used to specify the location, connect user, and configure number of threads to be submitted to Alchemi manager. The classes GridEncryptThread and GridDecryptThred are the thread classes that run under Alchemi and they use the DES algorithm.

The flow of GridCryptoGraphy program starts by dividing the raw file into several blocks and separating these blocks in order to parallelize the encryption process. The block separation process is done by reading the data file sequentially according to the block size. Each part of the file (block) is assigned to a thread including the last block whose size is the remainder of the predefined block size. The manager node passes the threads to the executer nodes. After the threads return with the encrypted results, GridCryptoGraphy saves the encrypted data to an output file according to the order of the threads.

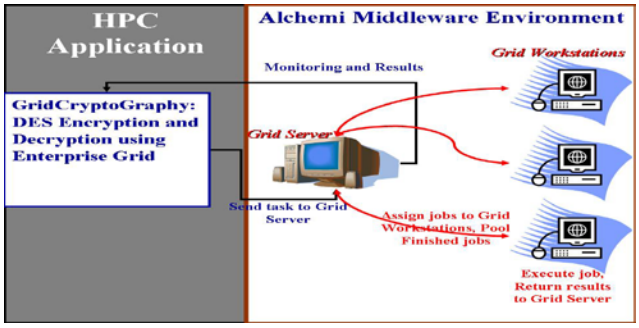


Fig.4: Symmetric Key Cryptography using Alchemi Middleware

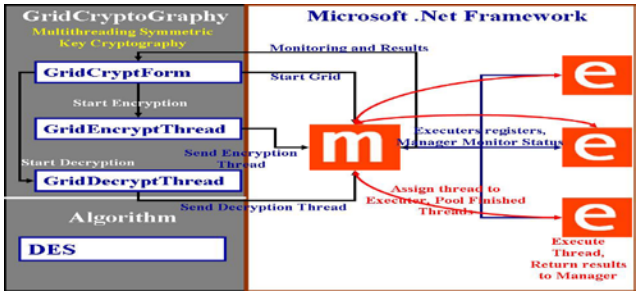


Fig. 5 GridCryptoGraphy Architecture

5. Cryptography using asynchronous data exchange

Our approach modeifies GridCryptoGraphy application to enable every executer accesses the input data file directly and at random by developing two new classes in this application. The classes are (AssGridEncryptThread and AssGridDecryptThread) which are the thread classes that run under Alchemi. The GridCryptoGraphy application only passes the names of input and output files to the manager as strings. It also passes the block size that each thread will access. So each thread will access the random access file according to each thread id multiplied with the block size, so if the block size is 1000, then thread number (0) will access the file in byte number (0) and thread

number (1) will access the file in byte number (1000), and so on. This process is accomplished through the reading and writing of the file, so the whole process is asynchronous as shown in Fig. 6. The job of the manager will only be initiating the threads, and not passing large datasets, thereby, we avoid delays due to creating large data blocks and passing them to executers.

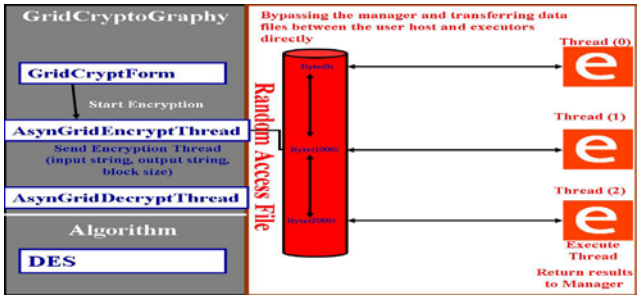


Fig. 6 GridCryptoGraphy Second experiment Architecture

6. Performance evaluation

Two experiments on large datasets have been conducted. Experiment 1 is a repetition of the GridCryptoGraphy application in [5] for a quantitative comparison with the results of experiment 2 that implements our approach. Eight executers have been used in each experiment with the following specification: for experiment 1, Intel® Pentium® 4 CPU 2.40 GHz, 512 MB RAM. For experiment 2, Intel® Pentium® 4 CPU 1.60 GHz, 128 MB RAM. Microsoft Windows XP Professional Version 2002 Service Pack 2. The nodes were interconnected over a shared LAN network of 100 Mbps.

The Alchemi manager was installed on a separate computer together with SQL Server 2000 and has the following specification: Intel® Pentium®4 CPU 3.00 GHz, 512 MB RAM. Microsoft Windows Server 2003 Standard Edition. The executions of the GridCryptoGraphy application run on the same computer with the manager.

A separate computer is used for monitoring the performance of the application with the following specification: Intel® Pentium®III CPU 731 MHz, 128 MB RAM. Microsoft Windows XP Professional Version 2002 Service Pack 2.

The encryption and decryption experiments were conducted on files of size 9645200 bytes (approximately 10 MB), 56610116 bytes (approximately 57 MB), 104858112 bytes (approximately 105 MB), 597393408 bytes (approximately 598 MB) and 1060842110 bytes (approximately 1061 MB) with different block sizes. For each file the encryption and decryption was carried on 1,2,3,4,5,6,7 and 8 executer nodes. The encryption experiments were conducted on file of size 104858112 bytes (approximately 105 MB) with 1, 5 and 10 Mb block size, which lead to the creation of 105, 21 and 11 work units respectively. For each experiment, the encryption was carried on 1, 2, 3, 4,5,6,7 and 8 executor nodes. Some snapshots of the program running are illustrated in Fig. 7(a-d).

The time performance results of experiment 2 are shown in Table 1.a and in Fig. 8.a. The speedup performance results are shown in Table 1.b and in Fig. 8.b where the speedup calculation is based on the following formula:
*Speedup=(((Time taken by 1 executer using 1 megabytes block size - Time taken by m executer using n megabytes block size)/ Time taken by 1 executer using 1 megabytes block size)*100)+100)*

Figure 9 and figure 10 show the comparison between the results of experiment 1 and experiment 2. Although executers used in experiment 1 have higher specifications (Pentium IV 2400 MHz processor and 512 MB of memory) than those (Pentium IV 1600 MHz processor and 128 MB of memory) used in experiment 2, It has been found that:

- Fig. 9 and Fig. 10 show remarkable improvements in the performance of our approach (experiment 2) compared to that of the traditional approach (experiment 1).
- In the first experiment, there is a drop in the performance after using 4 executers. In contrary, in the second experiment there was improvement in performance till 8 executers, therefore, larger files as the video file of size 1060842110 bytes (approximately 1061 MB) could successfully be encrypted.
- Although increasing the block size creates less work units and so the performance should be increased. It is found that the performance in experiment 1 is reduced compared with experiment 2.

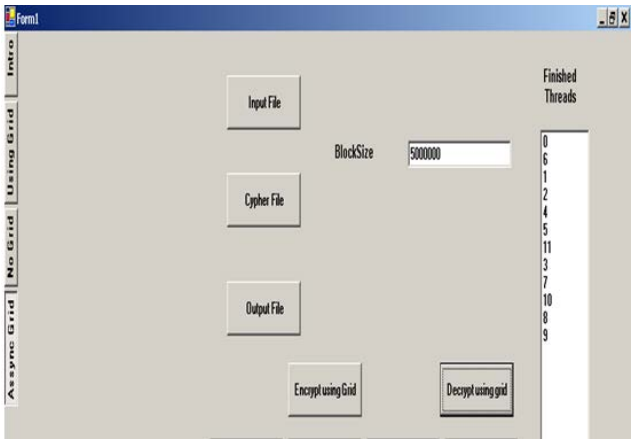


Fig. 7a: GridCryptoGraphy at runtime (monitoring of finished threads)

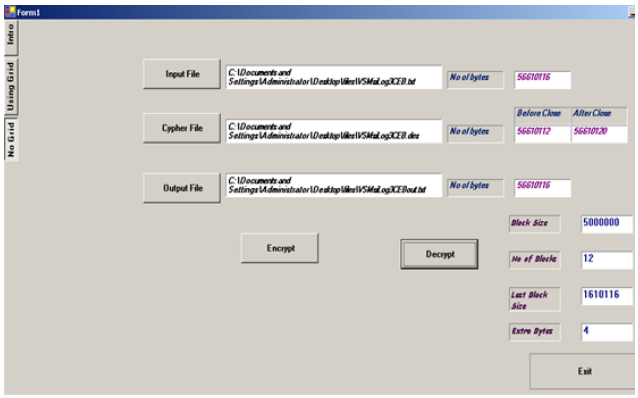


Fig.7b: GridCryptoGraphy at runtime (initializing files using 5-mega block size and 12 working unit)

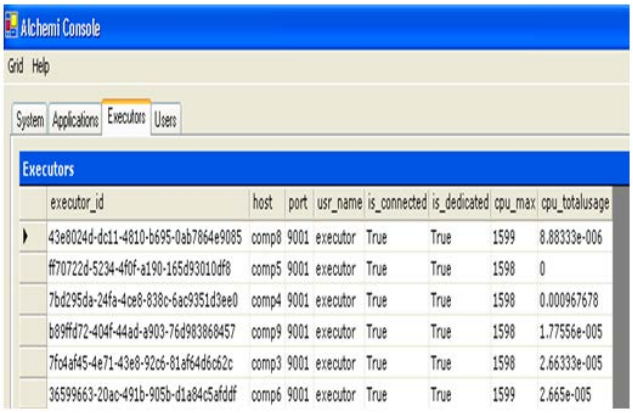


Fig. 7c: Six executers are working

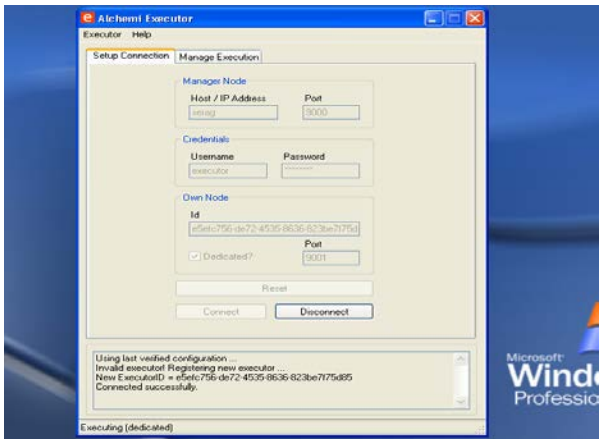


Fig. 7.d: Execution desktop

Table 1.a: Encryption time Performance results of 105Mega bytes file size

Block size / No of Executer	*(1 Mega) min :sec*	*(5 Mega) min :sec*	*(10 Mega) min: sec*
1	**00:42.563**	**00:35.469**	**00:35.141**
2	**00:23.625**	**00:26.328**	**00:24.063**
3	**00:23.469**	**00:24.266**	**00:24.013**
4	**00:22.641**	**00:23.328**	**00:23.375**
5	**00:21.859**	**00:21.281**	**00:23.078**
6	**00:20.078**	**00:20.828**	**00:22.188**
7	**00:21.313**	**00:20.391**	**00:20.172**
8	**00:21.547**	**00:18.469**	**00:20.141**

Table 1.b: Encryption Speedup Performance results of (105Mega bytes) file size

Block size / No of Executer	*(1 Mega) Speedup (%)*	*(5 Mega) Speedup (%)*	*(10 Mega) Speedup (%)*
1	**100.00**	**116.67**	**117.44**
2	**144.49**	**138.14**	**143.46**
3	**144.86**	**142.99**	**143.58**
4	**146.81**	**145.19**	**145.08**
5	**148.64**	**150.00**	**145.78**
6	**152.83**	**151.07**	**147.87**
7	**149.93**	**152.09**	**152.61**
8	**149.38**	**156.61**	**152.68**

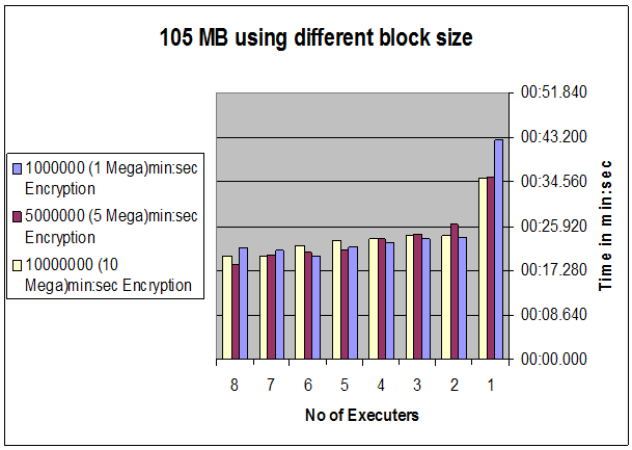


Fig. 8.a: Result graph of (104858112 bytes) file size with different block sizes

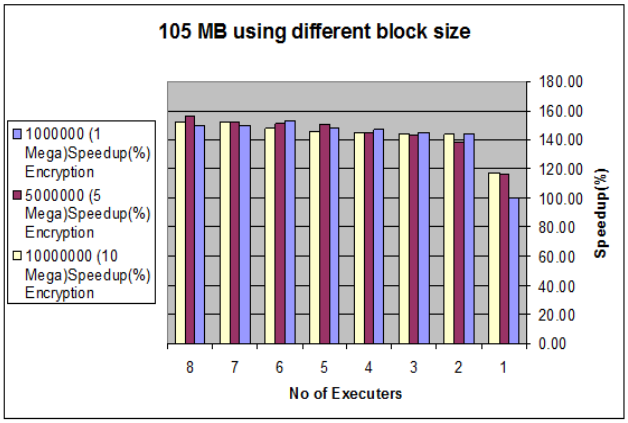


Fig. 8.b Speedup Result graph of (104858112 bytes) file size with different block sizes

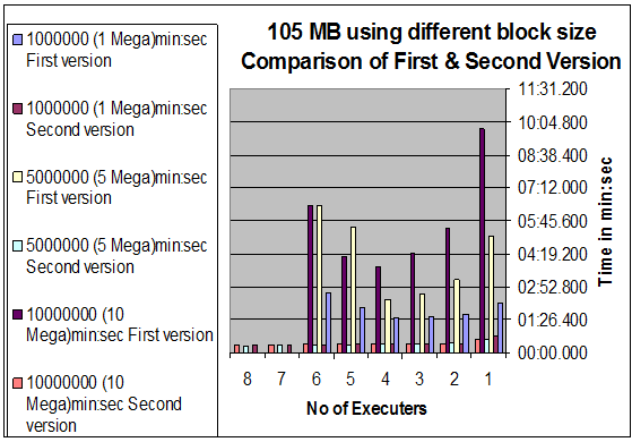


Fig 9.a: A time comparison of results to the first and second experiments (104858112bytes) file size with different block sizes

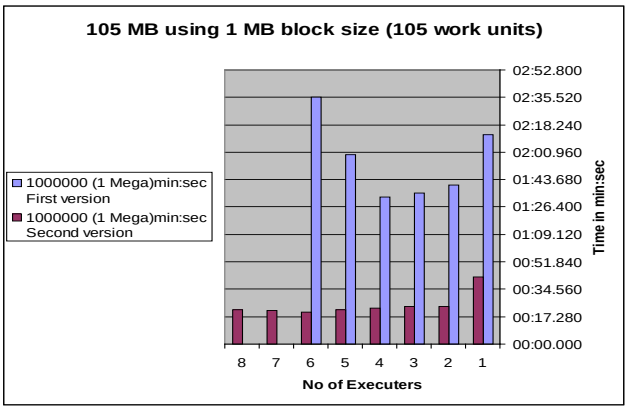


Fig. 9.b A time comparison of results to the First and Second Experiments of (104858112 bytes) file size with 1 Megabytes block sizes

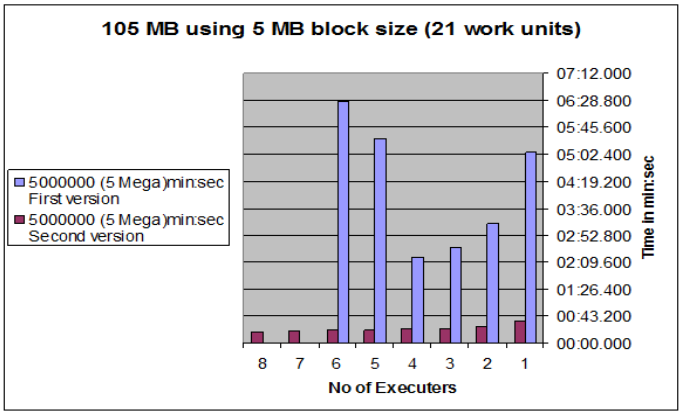


Fig. 9.c A time comparison of results to the First and Second Experiments of (104858112 bytes) file size with 5 Megabytes block sizes

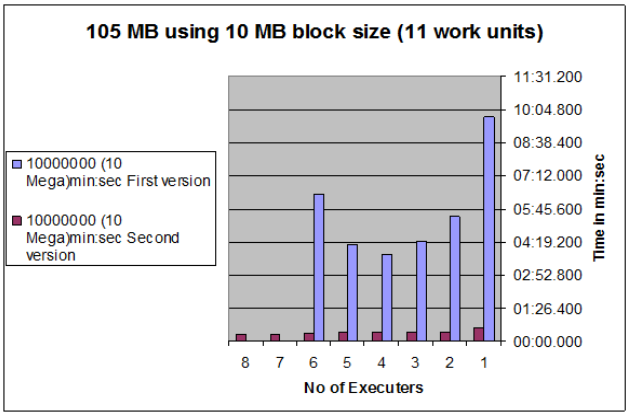


Fig. 9.d A time comparison of results to the First and Second Experiments of (104858112 bytes) file size with 10 Megabytes block sizes

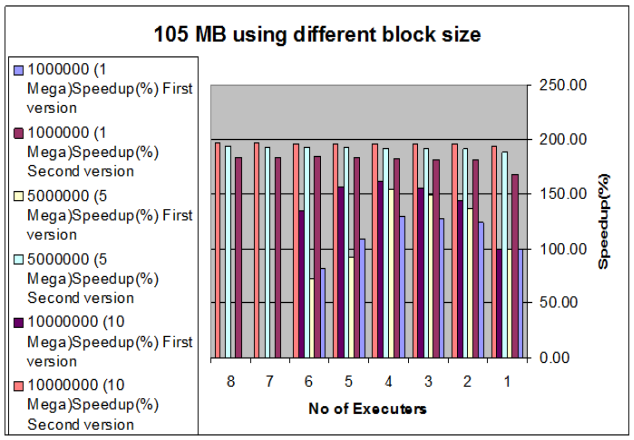


Fig. 10.a Speedup comparison of results to the First and Second Experiments of (104858112 bytes) file size with different block sizes

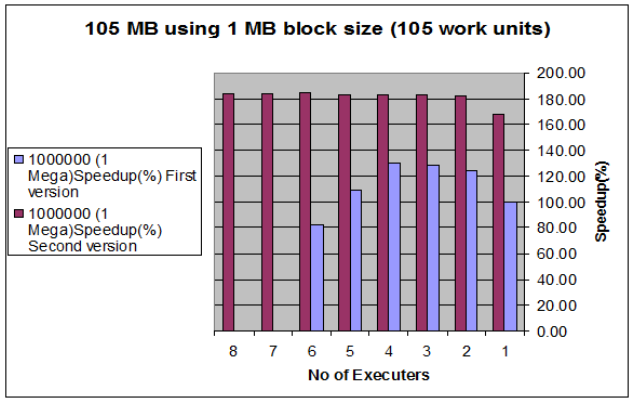


Fig (10.b) a speedup comparison of results to the First and Second Experiments of (104858112 bytes) file size with 1 Megabytes block sizes

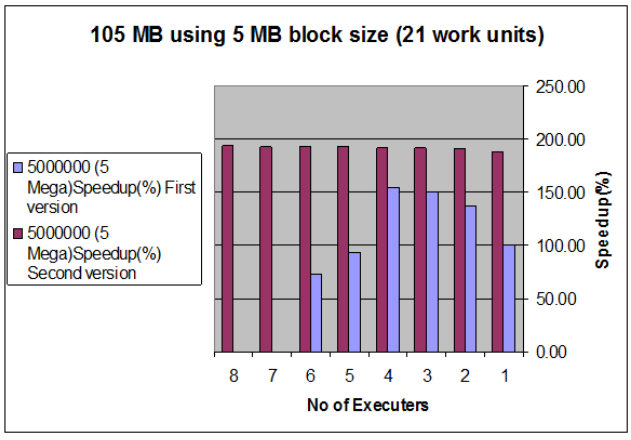


Fig. 10.c A speedup comparison of results to the First and Second Experiments of (104858112 bytes) file size with 5 Megabytes block sizes

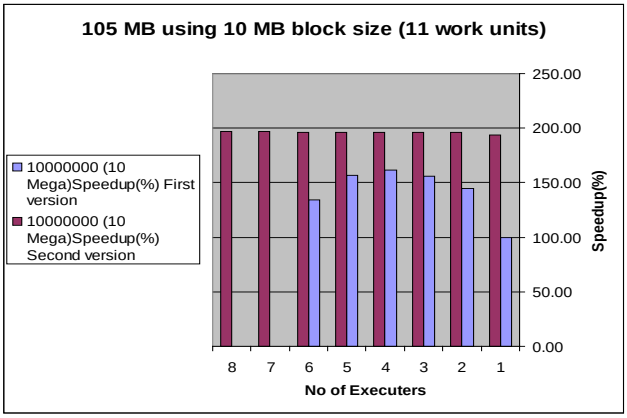


Fig (10.d) a speedup comparison of results to the First and Second Experiments of (104858112 bytes) file size with 10 Megabytes block sizes

7. Conclusions

This paper presents a grid based solution for solving the complex and large-scale problem of symmetric key cryptography that requires high performance computing resources. The problem was solved through a system-level middleware infrastructure called Alchemi. Alchemi is capable of creating an enterprise grid computing environment by harnessing windows machines and provide users with seamless computing ability and uniform access to resources in the heterogeneous grid environment. The proposed approach enhances the performance in terms of speed and limits the communication overhead. It is also scalable and cost-effective due to the effective and efficient utilization of a commodity-based high performance-computing platform.

References

[1] R. Buyya, D. Abramson and J. Giddy, "Driven Resource Management Architecture for Computational Power Grids". The 2000 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'2000), Las Vegas, 2000.

[2] http://recerca.ac.upc.edu/conferencies/AGC2007/, "Agent based Grid Computing".7th IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2007) Rio de Janeiro, Brazil, 154-17 May 2007.

[3] I. Foster and C. Kesselman, (editors), "The Grid: Blueprint for a New Computing Infrastructure", Morgan Kaufmann Publishers, USA, 1999.

[4] I. Foster, "Service-Oriented Science". Science, vol. 308, May 6, 2005.

[5] A. Setiawan, D. Adiutama, J. Liman, A. Luther and R. Buyya, "GridCrypt: High Performance Symmetric Key Cryptography Using Enterprise Grids". Liew, K. M. (editors) PDCAT, Springer-Verlag, pp. 872-877, 2004.

[6] I. Foster, "What is the Grid? A Three Point Checklist", GRIDToday, July 20, 2002.
[7] I. Foster, "The Grid: A New Infrastructure for 21st Century Science". Physics Today, 55(2):42-47, 2002.
[8] International Human Genome Sequencing Consortium. 2004
[9]The Human Proteome Folding Project, http://www. Grid.org/projects/hpf/
[10] http:// www.worldcommunitygrid.org/
[11] D. Kusnetzky and C. W. Olofson, "Oracle 10g: Putting Grids to Work", http://www.sswug.org/articles/viewarticle.aspx?id=18542
[12] http://setiathome.ss1.berkeley.edu/
[13] J. Bresnahan, M. Link, G. Khanna, Z. Imani, R. Kettimuthu and I. Foster. "Globus GridFTP: What's New in 2007" (Invited Paper), in Proceedings of the First International Conference on Networks for Grid Applications (GridNets 2007), Oct, 2007
[14] R. Kettimuthu, W. Allcock, L. Liming, J. Navarro and I. Foster. "GridCopy: Moving Data Fast on the Grid", in Proceedings of the Fourth High Performance Grid Computing Workshop to be held in conjunction with International Parallel and Distributed Processing Symposium (IPDPS 2007), March, 2007
[15] Legion – http://legion.verginia.edu/
[16] NASA IPG-http://www.ipg.nasa.gov
[17] Condor – http://www.cs.wisc.edu/condor/
[18] R. Ranjan, X. Chu, C. A. Queiroz, A. Harwood, R. Buyya. "A self organizing federation of Alchemi Desktop grids". Grids lab and P2P group, Australia, 2007.
[19] A. Luther, R. Buyya, R. Ranjan and S. Venugopal, "Alchemi: A .NET-based Grid Computing Framework and its Integration into Global Grids", Technical Report, GRIDS-TR-2003-8, Grid Computing and Distributed Systems Laboratory, University of Melbourne, Australia, December 2003.

Medhat Awadallah is an assistant professor at Electrical and Computer Engineering Department, Sultan Qaboos University. He obtained his PhD from university of Cardiff, UK. MSc and BSc from Helwan university, Egypt. His research interest includes cloud computing, sensor networks, high performance computing and real time systems.

Ahmed Youssef is an assistant professor at college of computer and information sciences, King Saud University. He obtained his P.h.D. and M.Sc. in computer science and engineering from university of Connecticut, USA. M.Sc and B.Sc in electronics and communications engineering from Helawn university, Egypt. His research interest includes cloud computing, mobile computing, high performance computing and information security.

A Luenberger State Observer for Simultaneous Estimation of Speed and Rotor Resistance in sensorless Indirect Stator Flux Orientation Control of Induction Motor Drive

Mabrouk Jouili[1], Kamel Jarray[2], Yassine Koubaa[1] and Mohamed Boussak[3], Senior Member, IEEE

[1] Research Unit of Automatic Control (UCA), University of Sfax, National Engineering School of Sfax (ENIS), Sfax, B.P. 1173, 3038 Sfax, Tunisia

[2] Research Unit of Modeling, Analysis and Control of Systems (MACS), University of Gabès, National Engineering School of Gabès (ENIG), Gabès 6029, Tunisia

[3] Laboratoire des Sciences de l'Information et des Systèmes (LSIS), University of Aix Marseille III, Ecole Centrale de Marseille (ECM), Marseille, UMR 6168, France

Abstract

The primary objective of this paper is to implement a sensorless indirect stator field oriented control (ISFOC) of induction motor drive with rotor resistance tuning. Indeed, the proposed method for simultaneous rotor speed and rotor resistance estimation is based on Luenberger observer (LO). In order to estimate the rotor speed and the rotor resistance, an adaptive algorithm based on Lyapunov stability theory by using measured and estimated stator currents and estimated stator flux is proposed. The suggested control scheme, as a result, achieves a sound performance with computational complexity reduction obtained by using the analytical relation to determine the LO gain matrix. Again, the observer is simple and robust, when compared with the previously developed observers, and suitable for online implementation. For current regulation, however, this paper suggests a conventional Proportional-Integral (PI) controller with feed-forward compensation terms in the synchronous frame. Owing to its advantages, an Integral-Proportional (IP) controller is used for rotor speed regulation. The design, analysis, and implementation for a 3-kW induction motor are completely carried out using a dSpace DS 1104 digital signal processor (DSP) based real-time data acquisition control (DAC) system, and MATLAB/Simulink environment. Digital simulation and experimental results are presented to show the improvement in performance of the proposed algorithm.

Keywords: *Stator flux orientation control (ISFOC), Sensorless vector control, rotor resistance estimation, feedforward decoupling, induction motor drive, Luenberger state-observer (LSO).*

1. Introduction

Due to their high performances in terms of reliability, robustness and efficiency, the adjustable ac-motor drives are increasingly adopted in most industrial applications such as military, aerospace and automotive industries [1]. The aforementioned qualities have been researched and improved by using intelligent and sophisticated control methods based on the field oriented control (FOC) because it adjusts both the amplitude and phase of ac-excitations. The vector control technique has been widely used for high-performance induction motor drives where the knowledge of the rotor speed is necessary. This information is provided by an incremental encoder, which is the most common positioning transducer used today in industrial applications [2, 3].

The use of this sensor implies more electronics, higher cost, lower reliability, difficulty in mounting in some cases such as motor drives in harsh environment and high speed drives, increase in weight, increase in size, and increase in electrical susceptibility.

To overcome these problems, in recent years, the elimination of the transducers has been considered as an attractive prospect. Therefore, numerous approaches have been proposed to estimate the rotor velocity and/or position.

In hottest literature, many researchers have carried out the design of sensorless vector control induction motor drives. These methods, definitely, are based on the following schemes:

Model Reference Adaptive System (MRAS): [4, 7].

Extended Kalman Filters (EKF): [8, 10].

Extended Luenberger observer (ELO): [12, 17].

Newly fuzzy logic and neuronal networks observers [18, 19].

Indeed, some of these methods require specially modified machines and the injection of disturbance signals or the use of a machine model. Otherwise, all other methods for speed estimation using a machine model fed by stator quantities are parameter dependent; therefore, parameter errors can degrade the speed control performance. Thus, some kind of parameter adaptation is required in order to obtain high-performance sensorless vector control drive. At very low speed, indirect stator-flux-oriented control (ISFOC) of induction motor drive is particularly sensitive to an accurate rotor resistance value in the stator flux. To prevail over this problem, online tuning rotor resistance variation of the induction motor is essential in order to maintain the dynamic performance of a sensorless ISFOC drive. Recently, many works dealing with drives without shaft transducers have been developed using different approaches to estimate rotor speed and rotor resistance [11, 20]. Most of these approaches require additional sensors that were not strictly used in standard ISFOC drive; thus, increasing cost and complexity may rule out practical use.

In the very paper, we suggest a contribution to the issue of sensorless indirect stator-flux-oriented control (ISFOC) of IM drive with rotor resistance tuning. The rotor speed and rotor resistance is estimated by the designed-observer by relying on the measured and estimated stator currents and estimated stator fluxes. As a matter of fact, this observer is designed to simultaneously estimate the rotor speed, the rotor resistance, the stator flux and the stator currents. In this respect, the singular perturbation theory is used to get a sequential and simple design of the observer, and the flux observer stability is ensured through the Lyapunov theory. Afterwards, a full description and justification of the proposed algorithm is given and its performances are tested by simulations and experiments. Although related algorithms have previously been presented, the following contributions are believed to be new. First, the dynamic and steady-state performances are analyzed. Excellent behaviour is verified in most cases. Second, the use of the stator field oriented control and a general framework is developed.

This paper is organized as follows: in Section 2, we briefly review the indirect stator-flux-oriented control (ISFOC) of induction motor drives. The procedure design proposed to simultaneous rotor speed and rotor resistance estimation is described in Section 3. Experimental and simulation results are presented in section 4. Finally, in Section 5 we give some comments and conclusions.

2. Stator Flux Orientation Strategy

For ISFOC, the stator flux vector is aligned with d-axis and sets the stator flux to be constant equal to the rated flux, which means $\Phi_{ds} = \Phi_s$ and $\Phi_{qs} = 0$.

2.1 Induction Motor Model

The dynamic model of an induction motor can be represented according to the usual d-axis and q-axis components in synchronous rotating frame as

$$v_{ds} = \frac{R_S(\tau_S+\tau_r)}{\tau_r}\left(1+\frac{\sigma\tau_S\tau_r}{\tau_S+\tau_r}p\right)i_{ds} - \sigma L_s\omega_{sl}i_{qs} - \frac{\Phi_S}{\tau_r} \quad (1)$$

$$v_{ds} = \frac{R_S(\tau_S+\tau_r)}{\tau_r}\left(1+\frac{\sigma\tau_S\tau_r}{\tau_S+\tau_r}p\right)i_{qs} + \sigma L_s\omega_{sl}i_{ds} + \omega_r\Phi_s \quad (2)$$

$$\Phi_s = Ls\frac{1+\sigma\tau_r p}{1+\tau_r p}i_{ds} - \frac{\sigma\tau_r L_S\omega_{sl}}{1+\tau_r p}i_{qs} \quad (3)$$

$$\omega_{sl} = \frac{Ls}{\tau_r}\frac{1+\sigma\tau_r p}{\Phi_S-\sigma L_S i_{ds}}i_{qs} \quad (4)$$

$$T_e = n_p\Phi_s i_{qs} \quad (5)$$

Where $\omega_{sl} = \omega_s - \omega_r$; $\tau_r = \frac{L_r}{R_r}$; $\tau_S = \frac{L_S}{R_S}$; and $\sigma = 1 - \frac{M^2}{L_S L_r}$

It can be perceived that if the stator flux is kept constant, the torque can be controlled by the q-axis current.

The electromagnetic torque equation and the electrical angular speed motor are related by:

$$J\frac{d\omega_r}{dt} + f\omega_r = n_p(T_e - T_l) \quad (6)$$

2.2 Feedforward decoupling controller

It can be seen that the d-axis and q-axis voltage equations are coupled by the terms $-\sigma L_s\omega_{sl}i_{qs} - \frac{\Phi_S}{\tau_r}$ and $\sigma L_s\omega_{sl}i_{ds} + \omega_r\Phi_S$. These terms are considered as disturbances and are cancelled by using a decoupling method that utilizes non-linear feedback of the coupling voltages. If the decoupling method is implemented, the voltage equations become [18].

$$v_d = v_{ds} + e_d = \frac{R_S(\tau_S+\tau_r)}{\tau_r}\left(1+\frac{\sigma\tau_S\tau_r}{\tau_S+\tau_r}p\right)i_{ds} \quad (7)$$

$$v_q = v_{qs} + e_q = \frac{R_S(\tau_S+\tau_r)}{\tau_r}\left(1+\frac{\sigma\tau_S\tau_r}{\tau_S+\tau_r}p\right)i_{qs} \quad (8)$$

Where $e_d = \sigma L_s\omega_{sl}i_{qs} + \frac{\Phi_S}{\tau_r}$ and $e_q = -\sigma L_s\omega_{sl}i_{ds} - \omega_r\Phi_S$; e_d and e_q are, respectively, the d-axis and q-back electromotive force (EMF).

Hence, the dynamics of the d-axis and q-axis currents are now represented by simple linear first-order differential equations. Therefore, it is possible to effectively control the currents with a PI controller.

In Fig. 1, k_{ii} and k_{ip} denote the proportional and integral gains of the PI d, q axis current controller, respectively.

$G_d(p)$, $G_q(p)$ are the no decoupling electrical d, q axis transfer functions of the induction machine. It should be noted that the estimated values, denoted by ^, are introduced to cancel out the coupling terms in the induction motor model.

If we assume that the back EMFs are canceled by the feedforward compensation term, we obtain

$$G_d(p) = G_q(p) = \frac{K_c}{1+\tau_c p} \tag{9}$$

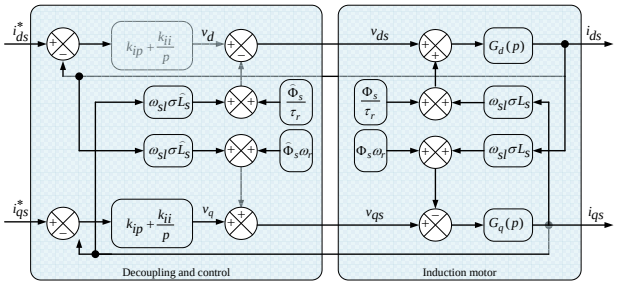


Figure. 1 Block diagram of the conventional PI controller with feed forward decoupling method.

Where $K_c = \frac{\tau_r}{R_s(\tau_s+\tau_r)}$ is a gain and $\tau_c = \frac{\sigma\tau_r\tau_s}{\tau_s+\tau_r}$ is a time constant.

The closed-loop current transfer function is

$$\frac{i_{ds}(p)}{i^*_{ds}(p)} = \frac{i_{qs}(p)}{i^*_{qs}(p)} = \frac{\omega_n^2}{p^2+2\xi\omega_n p+\omega_n^2}\left(1+\frac{k_{ip}}{k_{ii}}p\right) \tag{10}$$

This allows us to write

$$\begin{cases} k_{ii} = \dfrac{\tau_c\omega_n^2}{K_c} \\ k_{ip} = \dfrac{2\xi\tau_c\omega_n - 1}{K_c} \end{cases} \tag{11}$$

Where $\omega_n = \sqrt{\frac{K_c k_{ii}}{\tau_c}}$ and $\xi = \frac{1+K_c+k_{ip}}{2\omega_n\tau_c}$.

ω_n and ξ indicate the natural frequency and damping ratio, respectively. When the dynamics of the d- and q-axes currents are equivalent, the PI gains can be copied to the q-axis controller.

3. Adaptive Luenberger Observer

3.1 Flux observer of induction motor

The state model of the induction motor can be described in a rotating reference frame by:

$$\begin{cases} \dot{x}(t) = \mathbf{A}x(t) + \mathbf{B}v_s(t) \\ y(t) = \mathbf{C}x(t) \end{cases} \tag{12}$$

The adaptive full observer for the estimation of the stator current and the stator flux, using the measured stator currents and voltages, is described by the following set of equations:

$$\begin{cases} \dot{\hat{x}}(t) = \hat{\mathbf{A}}\hat{x}(t) + \mathbf{B}v_s(t) + \mathbf{L}\left[y(t) - \mathbf{C}\hat{x}(t)\right] \\ \hat{y}(t) = \mathbf{C}\hat{x}(t) \end{cases} \tag{13}$$

Where : $\bar{x} = \begin{bmatrix} \bar{i}_{ds} & \bar{i}_{qs} & \bar{\Phi}_{ds} & \bar{\Phi}_{qs} \end{bmatrix}^T$; $x = \begin{bmatrix} i_{ds} & i_{qs} & \Phi_{ds} & \Phi_{qs} \end{bmatrix}^T$;

$\bar{y} = \begin{bmatrix} \bar{i}_{ds} & \bar{i}_{qs} \end{bmatrix}^T$; $y = \begin{bmatrix} i_{ds} & i_{qs} \end{bmatrix}^T$; $v_s(t) = \begin{bmatrix} v_{ds} & v_{qs} \end{bmatrix}^T$

$$\mathbf{A} = \begin{bmatrix} -\gamma\mathbf{I_2} + \omega_r\mathbf{J} & \dfrac{1}{\sigma L_s\tau_r}\mathbf{I_2} - \dfrac{1}{\sigma L_s}\omega_r\mathbf{J} \\ -R_s\mathbf{I_2} & \mathbf{O_2} \end{bmatrix}; \quad \mathbf{B} = \begin{bmatrix} \dfrac{1}{\sigma L_s}\mathbf{I_2} \\ \mathbf{I_2} \end{bmatrix};$$

$\mathbf{C} = \begin{bmatrix} \mathbf{I_2} & \mathbf{O_2} \end{bmatrix}$; $\mathbf{I_2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$; $\mathbf{J} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$; $\mathbf{O_2} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ and

$\gamma = \frac{1}{\sigma}\left(\frac{\tau_s+\tau_r}{\tau_s\tau_r}\right)$.

Where ^ alludes to the estimated values, $\hat{x}(t)$ is the observer state vector and $\mathbf{L}$ is the observer gain matrix which is selected so that the system will be stable.

3.2 Luenberger Observer Gain Design

To ensure that the estimation error vanishes over time for any $\tilde{x}(0)$, we should select the observer gain matrix $\mathbf{L}$ so that $(\mathbf{A}-\mathbf{LC})$ is asymptotically stable. Consequently, the observer gain matrix should be chosen so that all eigenvalues of $(\mathbf{A}-\mathbf{LC})$ have real negative parts.

To ensure stability for all ranges of speed, the conventional procedure is to select the observer poles proportional to the motor poles (the proportionality constant is $kp \succ 1$). If the poles of the induction motor are given by λ_{IM} , the observer poles λ_{LO} are selected as:

$$\lambda_{LO} = k_p\lambda_{IM} \tag{14}$$

This can be achieved by defining the observer matrix $\mathbf{L}$ in a special form

$$\mathbf{L} = \begin{bmatrix} l_1\mathbf{I_2} + l_2\mathbf{J} \\ l_3\mathbf{I_2} + l_4\mathbf{J} \end{bmatrix} = \begin{bmatrix} \mathbf{L_1} \\ \mathbf{L_2} \end{bmatrix} \tag{15}$$

To determine the eigenvalues of the matrix $\mathbf{A}$, we use:

$$\det(\lambda_{IM}\mathbf{I_2} - \mathbf{A}) = \begin{vmatrix} \lambda_{IM} + \gamma\mathbf{I_2} - \omega_r\mathbf{J} & -\frac{1}{\sigma L_s \tau_r}\mathbf{I_2} + \frac{1}{\sigma L_s}\mathbf{J} \\ R_s\mathbf{I_2} & \lambda_{IM} \end{vmatrix} = 0 \quad (16)$$

In order to simplify the equation we define

$$a = \gamma\mathbf{I_2} - \omega_r\mathbf{J} \;;\; b = -\frac{1}{\sigma L_s \tau_r}\mathbf{I_2} + \frac{1}{\sigma L_s}\omega_r\mathbf{J} \text{ et } c = R_s\mathbf{I_2}$$

The characteristic equation of the matrix A is then

$$\lambda_{IM}^2 + a\lambda_{IM} - bc = 0 \quad (17)$$

To determine the eigenvalues of the matrix $(\mathbf{A}-\mathbf{LC})$

$$\det(\lambda_{LO}\mathbf{I_2} - (\mathbf{A}-\mathbf{LC})) = \begin{vmatrix} \lambda_{IM} + \gamma\mathbf{I_2} - \omega_r\mathbf{J} + \mathbf{L_1} & -\frac{1}{\sigma L_s \tau_r}\mathbf{I_2} + \frac{1}{\sigma L_s}\mathbf{J} \\ R_s\mathbf{I_2} + \mathbf{L_2} & \lambda_{LO} \end{vmatrix} = 0 \quad (18)$$

Hence, the characteristic equation is

$$\lambda_{LO}^2 + \lambda_{LO}(\mathbf{L_1} + a) - b(\mathbf{L_2} + c) = 0 \quad (19)$$

The substitution of Eq. (16) in (21) yields

$$k_p^2\lambda_{IM}^2 + k_p\lambda_{IM}(\mathbf{L_1} + a) - b(\mathbf{L_2} + c) = 0 \quad (20)$$

The identification of Eq. (22) and k_p^2* (19) gives the following results:

$$\begin{cases} l_1 = (k_p - 1)\gamma \\ l_2 = (k_p - 1)\hat{\omega}_r \\ l_3 = (k_p^2 - 1)R_s \\ l_4 = 0 \end{cases} \quad (21)$$

3.3 Adaptive Flux Observer for Speed Estimation

When the motor speed is not measured, it is treated as an unknown parameter in the observer (13). By adding an adaptive scheme for estimating the rotor speed to the observer, both states and unknown parameters can be estimated simultaneously. The adaptive scheme is derived using Lyapunov theory. From (12) and (13), the estimation error of the stator and rotor flux is given by the following equation:

$$\dot{e} = (\mathbf{A} + \mathbf{LC})e + (\quad)x \quad (22)$$

Where $e = x - \hat{x}$; $\Delta\mathbf{A} = \mathbf{A} - \hat{\mathbf{A}} = \begin{bmatrix} \Delta\omega_r\mathbf{J} & -\frac{1}{\sigma L_s}\Delta\omega_r\mathbf{J} \\ 0 & 0 \end{bmatrix}$ and

$\Delta\omega_r = \omega_r - \hat{\omega}_r$

We define a Lyapunov function candidate V

$$V = e_n^T e_n + \frac{(\hat{\omega}_r - \omega_r)^2}{\lambda} \quad (23)$$

Where λ is a positive constant and

$e_n = \begin{bmatrix} i_S - \hat{i}_S & \Phi_S - \hat{\Phi}_S \end{bmatrix} = \begin{bmatrix} e_{i_S} & e_{\Phi_S} \end{bmatrix} = \mathbf{\Gamma}e$

With $\mathbf{\Gamma}$ a nonsingular matrix

For the derivation of the adaptive mechanism, the unknown parameter $\hat{\omega}_r$ is considered constant. The time derivative of V becomes

$$\dot{V} = e_n^T\left\{\left[\mathbf{\Gamma}\left(\mathbf{A}+\mathbf{LC}\right)\mathbf{\Gamma}^{-1}\right]^T + \left[\mathbf{\Gamma}\left(\mathbf{A}+\mathbf{LC}\right)\mathbf{\Gamma}^{-1}\right]\right\}e_n + 2\frac{\Delta\omega_r}{\sigma L_S}(\hat{\Phi}_{\beta s}e_{i_{\alpha s}} - \hat{\Phi}_{\alpha s}e_{i_{\beta s}}) - \frac{2}{\lambda}\Delta\omega_r\frac{d\hat{\omega}_r}{dt} \quad (24)$$

The adaptive scheme for speed estimation is given by:

$$\hat{\omega}_r = K_{P\omega_r}\left(\hat{\Phi}_{\beta s}e_{\alpha s} - \hat{\Phi}_{\alpha s}e_{i_{\beta s}}\right) + K_{I\omega_r}\int_0^t\left(\hat{\Phi}_{\beta s}e_{i_{\alpha s}} - \hat{\Phi}_{\alpha s}e_{i_{\beta s}}\right)dt \quad (25)$$

The adaptive flux observer is stable according to the Lyapunov direct method if the observer gain is chosen such that the first term of (24) is negative semidefinite. This condition is fulfilled if the eigenvalues of $\mathbf{\Gamma}(\mathbf{A}+\mathbf{LC})\mathbf{\Gamma}^{-1}$ have negative real parts. Since the eigenvalues of $\mathbf{\Gamma}(\mathbf{A}+\mathbf{LC})\mathbf{\Gamma}^{-1}$ equal the eigenvalues of $(\mathbf{A}+\mathbf{LC})$ ($\mathbf{\Gamma}$ is nonsingular), the observer should have stable poles.

3.4 Adaptive Flux Observer for Speed and Rotor Resistance Estimation

Due to temperature changes during operation, the IM stator resistance and rotor resistance will vary. The proposed adaptive observer can be extended to include rotor resistance estimation. When both rotor resistance and speed are treated as unknown parameters of the observer, the estimation error of the stator and rotor flux is calculated from (12) and (13) and is provided by the following equation.

$$\dot{e} = (\mathbf{A} + \mathbf{LC})e + \Delta\mathbf{A}'x + \quad 'x \quad (26)$$

Where

$\Delta\mathbf{A}' = \begin{bmatrix} -\frac{\Delta R_r}{\sigma L_r}\mathbf{I_2} & \frac{\Delta R_r}{\sigma L_S L_r}\mathbf{I_2} \\ \mathbf{0_2} & \mathbf{0_2} \end{bmatrix}$ and $\Delta R_r = R_r - \hat{R}_r$

A Lyapunov function candidate V' is defined as follows:

$$V' = e_n^T e_n + \frac{(\hat{\omega}_r - \omega_r)^2}{\lambda} + \frac{(\hat{R}_r - R_r)^2}{\lambda'} = V + \frac{(\hat{R}_r - R_r)^2}{\lambda'} \quad (27)$$

and its time derivative is

$$\dot{V}' = \dot{V} + \hat{x}^T \mathbf{\Delta A}'^T \mathbf{\Gamma}^T e_n + e_n^T \mathbf{\Gamma \Delta A}' \hat{x}$$
$$-2\frac{\Delta R_r}{\sigma L_r L_s} e_{i\alpha s} (\hat{\Phi}_{\alpha s} - Ls\hat{i}_{\alpha s}) \quad (28)$$
$$-2\frac{\Delta R_r}{\sigma L_r L_s} e_{i\beta s} (\hat{\Phi}_{\beta s} - Ls\hat{i}_{\beta s}) + 2\frac{\Delta R_r}{\lambda'}\frac{d\hat{R}_r}{dt}$$

The adaptive scheme for rotor resistance estimation is found by equating the second and third term in (28)

$$\hat{R}_r = K_{PR_r}\left[e_{i\alpha s}(\hat{\Phi}_{\alpha s} - Ls\hat{i}_{\alpha s}) + e_{i\beta s}(\hat{\Phi}_{\beta s} - Ls\hat{i}_{\beta s})\right]$$
$$+K_{IR_r}\int_0^t \left[e_{i\alpha s}(\hat{\Phi}_{\alpha s} - Ls\hat{i}_{\alpha s}) + e_{i\beta s}(\hat{\Phi}_{\beta s} - Ls\hat{i}_{\beta s})\right]dt \quad (29)$$

The structure of the proposed adaptive observer for speed and rotor resistance estimation is shown in Fig. 2.

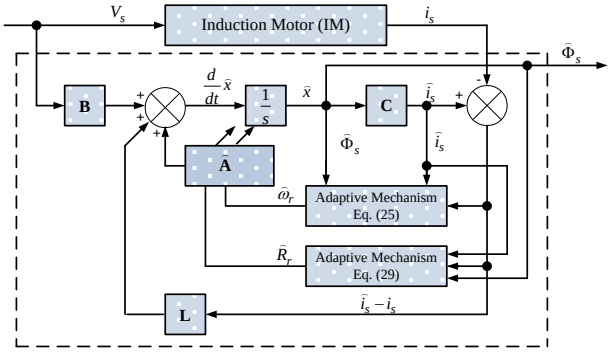


Figure. 2 Block diagram of combined speed and rotor resistance estimation.

4. Simulation and experimental results

In this section, some simulation and experimental results are presented to evaluate the effectiveness of the proposed control scheme for an induction motor. Figure 3 shows the block diagram of the proposed sensorless ISFOC with rotor resistance tuning of induction motor drive system. The bloc diagram consists of an induction motor, a PWM voltage source inverter, a field orientation algorithm, a coordinate translator, and a speed controller.

To implement the proposed sensorless ISFOC with rotor resistance tuning of induction motor drive, an experimental testing ground was carried out. It is essentially composed of:

- An induction motor a 3-kW whose parameters are listed in appendix.
- A static power electronics convertor from semikron composed of a diode rectifier and a three-leg voltage source IGBT inverter.
- Current sensors of Hall.
- A dSpace DS 1104 ACE Kit with control desk software plugged in a Pentuim 4 personal computer.

For the implementation of the proposed sensorless ISFOC ISFOC with rotor resistance tuning of induction motor drive, an experimental has been carried out (Fig. 4). The sensorless ISFOC algorithm which is programmed with Matlab-Simulink and downloaded in the dSpace 1104 control board offers a four-channel 16-bit (multiplexed) ADC and four 12-bit ADC units. A sampling period of 50µs is selected and the insulated gate bipolar transistors (IGBTs) are working at a switching frequency of 10 kHz with a dead time of 20µs. The load is generated through a magnetic powder brake coupled to the induction motor. The output control signals of the Slave I/O PWM are of TTL level 5V, whereas IGBTs of the static inverter must receive signals of 15V. Additionally, an adaptation interface board using the integrated circuit IR2130 from International rectifier is realized.

In order to check the validity of the implementation of ISFOC with rotor resistance tuning of induction motor drive using dSpace DS 1104 control board, some simulation and experimental works have been performed. The flux is kept constant at its rated value 1.21Wb. The first aim of the present simulation and experimental results is to test the performance of the sensorless ISFOC with rotor resistance tuning of induction motor drive system for a reference speed ±1000 rpm with load torque equal 20 N.m applied and removed at t = 6.5 and 16.5 s, respectively.

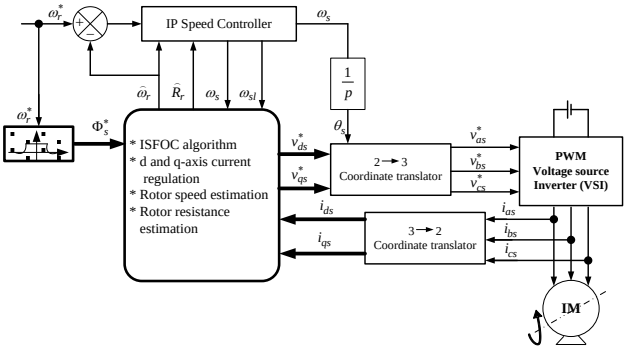


Figure.3 Block diagram of sensorless (ISFOC) with rotor resistance tuning of induction motor drive system.

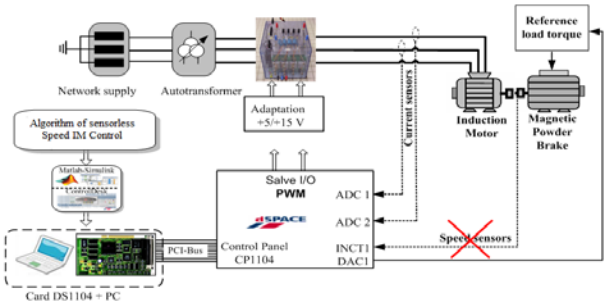
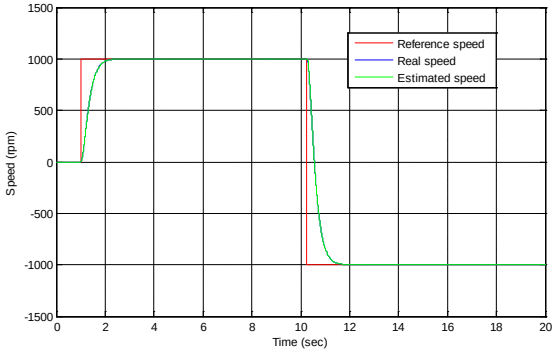


Figure.4 Scheme used for experimental setup.

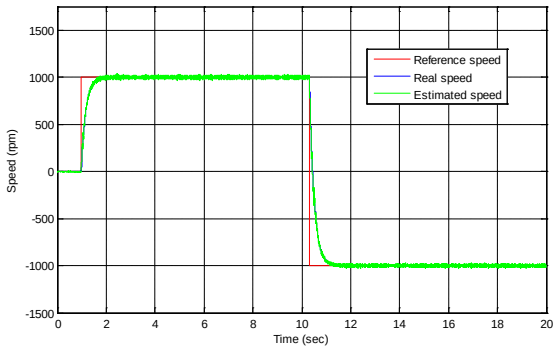
The IP speed controller is designed to stabilize the speed control loop. Moreover, the gains of the IP controller are determined by using a design method to obtain a damping ratio of 1.

As a first test, Fig. 5 shows the simulation and experimental results for sensorless (ISFOC) with rotor resistance tuning of induction motor drive. Accordingly, when the speed command changes from zero to (1000 rpm) in forward rotation, it changes to reverse direction of the same speed.

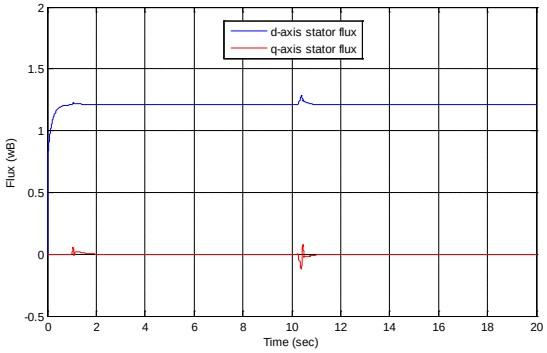
In Figs. 5(a) and (b), the simulation and experimental results of the reference, estimated and actual rotor speed, and in Fig. 5(c) and (d), simulation and estimated results of d-axis and q-axis stator flux, are presented. The estimated d, q components of stator flux are obtained from the stator voltage model of induction motor in d, q reference frame. Besides, simulation and measured results of d-q axis currents are given in Figs. 5(e) and (f). Fig. 5(g) and (h) shows, respectively, simulation results and experimental of estimated rotor resistance.



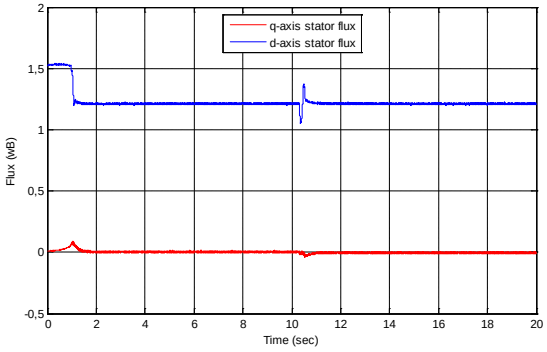
(a) simulated



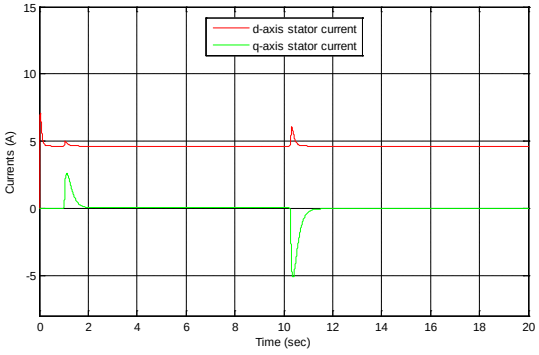
(b) experimental



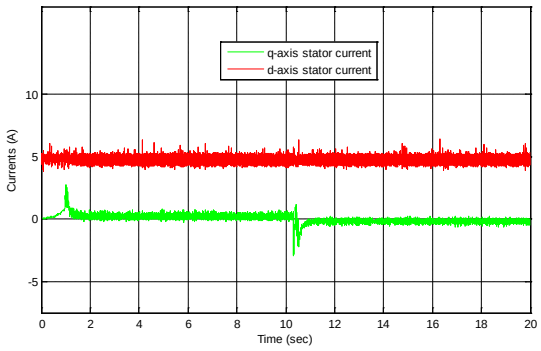
(c) simulated



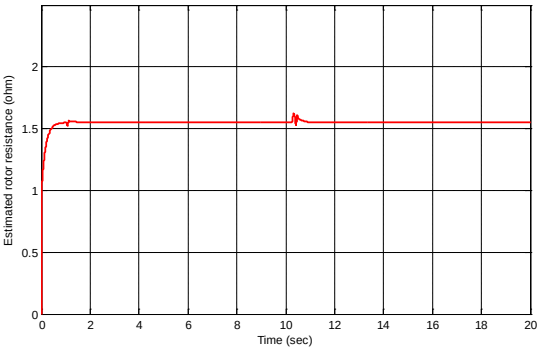
(d) experimental



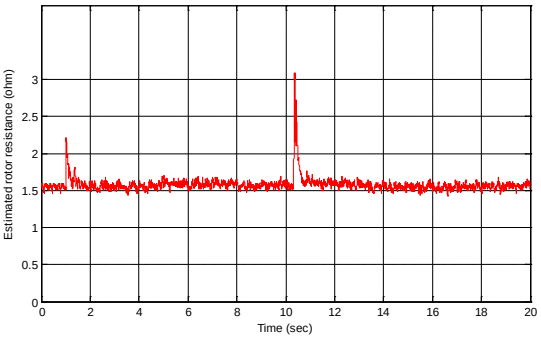
(e) simulated



(f) experimental



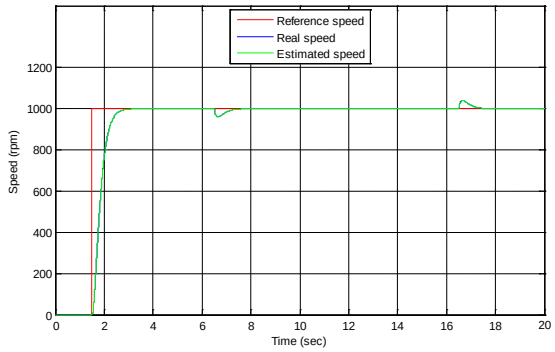
(g) simulated



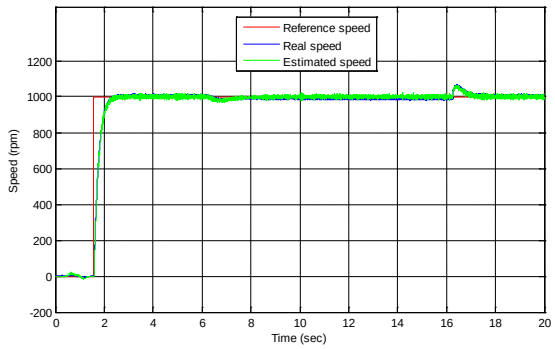
(h) experimental

Figure.5 Experimental and simulation results of sensorless (ISFOC) (no load) with rotor resistance tuning for reversing speed reference from 1000rpm to -1000rpm.

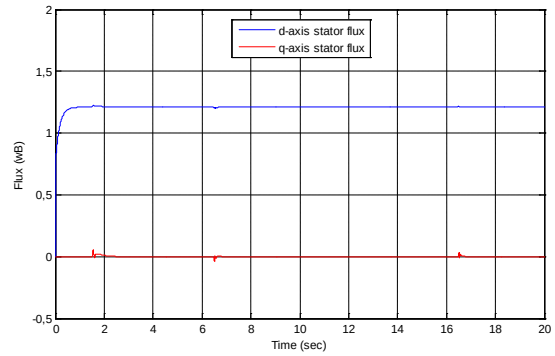
As a second test, some simulation and experimental results for sensorless (ISFOC) with rotor resistance tuning of induction motor drive are presented in Fig. 6 for 1000 rpm speed reference command and a load torque of 20 N.m is applied and removed at t=6.5 and 16.5s, respectively. In Figs. 6(a) and (b), the simulation and experimental results of the reference, real and estimated rotor speed, and in Figs. 6(c) and (d), the simulation and estimated results of d-axis and q-axis stator flux, are presented a load torque variation. In Figs. 6(e) and (f), the simulation and experimental results of d-q axis currents are presented. Furthermore, the simulation results and measured stator phase currents are given in Figs. 6(g) and (h). Figs. 6(i) and (j) show, respectively, the simulation results and measurement of load torque and q-axis stator current. Fig. 6(k) and (l) shows, respectively, the simulation and experimental results of estimated rotor resistance.



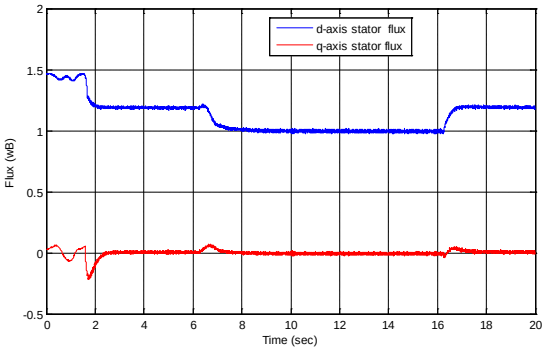
(a) simulated



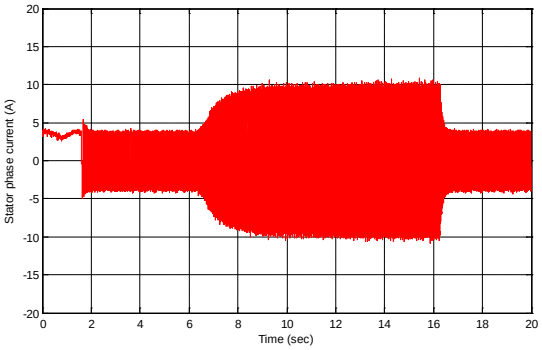
(b) experimental



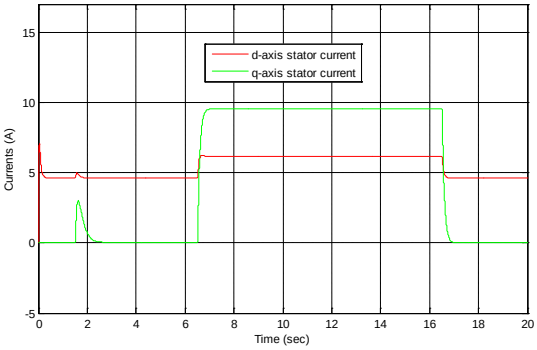
(c) simulated



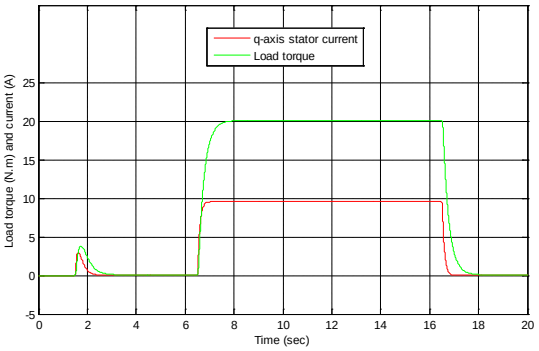
(d) Experimental



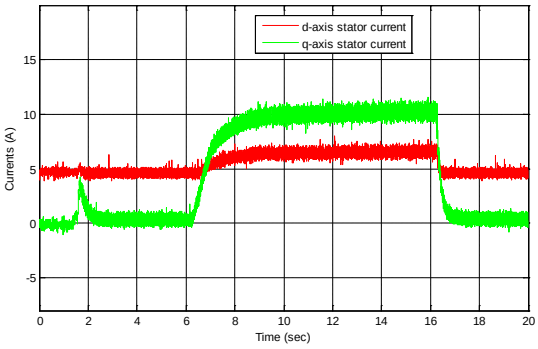
(h) Experimental



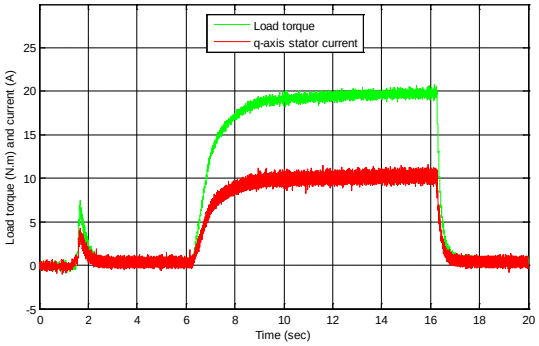
(e) simulated



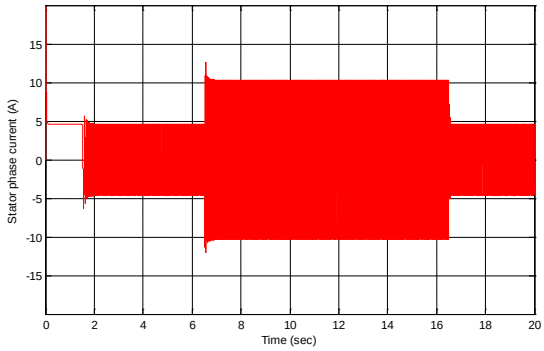
(i) simulated



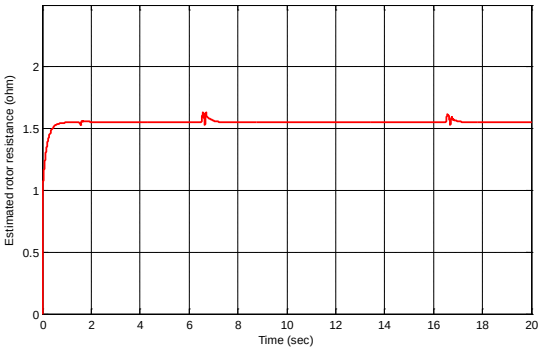
(f) Experimental



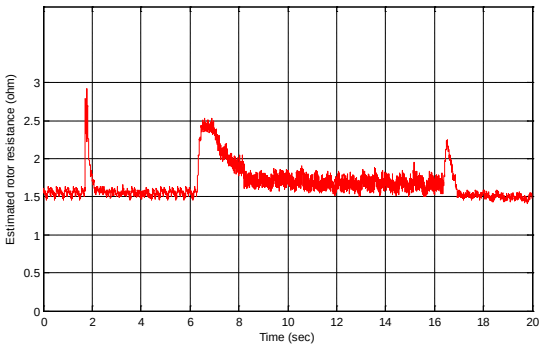
(j) Experimental



(g) simulated



(k) simulated



(l) Experimental

Figure.6 Experimental and simulation results of sensorless (ISFOC) (with load torque of 20 N.m applied at 6.5 s) with rotor resistance tuning the speed command is 1000 rpm

In steady state operation, it should be noted that in Fig. 6(c) and (d) the d-axis stator flux builds up to the rated value (1.21 wB) by d-axis stator current, while q-axis stator flux and current components remain zero. This shows that a decoupling between stator flux and the torque is achieved. It should also be noted that we have considered the stator resistance to be constant. However, like stator resistance, rotor resistance also depends on temperature. It is clear that an improvement of high performance sensorless speed control requires tracking changes in stator resistance.

5. Conclusions

In this paper, one has validated the online simultaneous estimation of speed and rotor resistance in sensorless indirect stator flux orientation control of induction motor drive system based on Luenberger observer. In other words, the complexity of the algorithm is reduced by using analytical relations to obtain directly the Luenberger observer (LO) gain matrix as a function of the electrical velocity and the proportional constant. The validity of the proposed sensorless ISFOC of induction motor drive with rotor resistance tuning was also proven by simulation and experiments for a wide range of speed. More importantly, all experimental results confirm the good dynamic performances of the developed drive systems and show the validity of the suggested method. It is concluded from the results presented in this paper that the proposed scheme performs well for both high and low speed.

Appendix

List of motor specification and parameters: 220/380V, 3kW, 4 poles, 1430 rpm

$R_s = 2.3\Omega$; $R_r = 1.55\Omega$; $L_s = L_r = 0.261H$; $M = 0.245H$; $f = 0.002 Nm.s.rd^{-1}$; $J = 0.03 kg.m^2$.

References

[1] J. Jung, K. Nam, "A dynamic decoupling control scheme for high speed operation of induction motor", IEEE Transaction on Industry Applications, Vol. 30, 1994, pp. 1219– 1224.

[2] F. J. Lin, R. J. Wai, C. H. Lin, and D. C. Liu, "Decoupling stator-flux oriented induction motor drive with fuzzy neural network uncertainly observer", IEEE Transaction on Industrial Electronics, Vol. 47, No. 2, 2000, pp. 356– 367.

[3] S. Suwankawin and S. Sangwongwanich, "A speed sensorless IM drive with decoupling control and stability analysis of speed estimation", IEEE Transaction on Industrial Electronics, Vol. 49, No. 2, 2002, pp. 444– 455.

[4] Y. Agrebi, M. Triki, Y. Koubaa, M. Boussak, "Rotor speed estimation for indirect stator flux oriented induction motor drive based on MRAS Scheme", Journal of Electrical Systems, 2007, Vol. 3, pp. 131– 143.

[5] R. Blasco-Giménez, G. Asher, M. Summer, K. Bradley, "Dynamic performance limitations for MRAS based sensorless induction motor drives. Part 1: Stability analysis for the closed loop drive", IEEE Proceeding of Electric Power Applications, Vol. 143, 1996, pp. 113– 122.

[6] Y. Koubaa, M. Boussak, "Rotor resistance tuning for indirect stator flux oriented induction motor drive based on MRAS scheme", Revue European Transactions on Electrical Power, Vol. 15, No. 6, 2005, pp. 557– 570.

[7] L. Zhen and L. Xu, "Sensorless field orientation control of induction machines based on mutual MRAS scheme", IEEE Transaction on Industry Applications, Vol. 45, No. 5, 1998, pp. 824– 831.

[8] Y. R. Kim, S. K. Sul, and M. H. Park, "Speed sensorless vector control of induction motor using an extended Kalman filter", IEEE Transaction on Industry Applications, Vol. 30, No. 5, 1994, pp. 1225– 1233.

[9] K. L. Shi, Y. K. Wong, and S. L. Ho, "Speed estimation of an induction motor drive using an optimized extended Kalman filter", IEEE Transaction on Industrial Electronics, Vol. 49, No. 1, 2002, pp. 124– 133.

[10] L.C. Zai, C.L. Demarco, T.A. Lipo, "An extended Kalman filter approach to rotor time-constant measurement in PWM induction motor drives", IEEE Transaction on Industrial Electronics, Vol. 28, No. 1, 1992, pp. 96– 104.

[11] K. Akatsu, A. Kawamura, "Online rotor resistance estimation using the transient state under the speed sensorless control of induction motor", IEEE Transaction on Power Electronics, Vol. 15, No. 3, 2000, pp. 553– 560.

[12] L. Jingchuan, L. Xu, Z. Zhang, "An adaptive sliding-mode observer for induction motor sensorless speed control", IEEE Transaction on Industrial Electronics, Vol. 46, No. 1, 1999, pp. 100– 110.

[13] H. Kubota K. Matsuse, "Speed sensorless field-oriented control of induction motor with rotor resistance adaptation", IEEE Transaction on Industrial Electronics, Vol. 30, No. 5, 1994, pp. 1219– 1224.

[14] H. Kubota, K. Matsuse, T. Nakano, "Dsp-based speed adaptive flux observer of induction motor", IEEE

Transaction on Industrial Electronics, Vol. 29, 1993, pp. 344– 348.
[15] D.J. Atkinson, J.W. Finch, P.P. Acamley, "Estimation of rotor resistance in induction motor", IEE Proceedings Electric Power Applications, Vol. 143, No. 1, 1996, pp. 87– 94.
[16] T. Du, P. Vas, F. Stronach, "Design and application of extended observers for joint state and parameter estimation in high-performance AC drivers", IEE Proceedings Electric Power Applications, Vol. 142, No. 2, 1995, pp. 71– 78.
[17] T.O Kowalska, "Application of extended Luenberger observer for flux and rotor time-constant estimation in induction motor drives", IEE Proceedings on Control Theory and Applications, Vol. 136, No. 6, 1989, pp. 324– 330.
[18] H. Rehman, R. Dhaouadi, A fuzzy learning-Sliding mode controller for direct field-oriented induction machines, Neurocomputing. 71 (2008) 2693– 2701.
[19] B. Karanayil, M. F. Rahman, and C. Grantham, "Online stator and rotor resistance estimation scheme using artificial neural networks for vector controlled speed sensorless induction motor drive", IEEE Transaction on Industrial Electronics, Vol. 54, No. 1, 2007, pp. 167– 176.
[20] M. Boussak, K. Jarray, "A High-Performance Sensorless Indirect Stator Flux Orientation of Induction Motor Drive", IEEE Transaction on Industrial Electronics, Vol. 53, No. 1, 2006, pp. 41– 49.

Mabrouk Jouili was born in Ben Guerdane, Tunisia, on August 14, 1980. He received the Engineer degree and the M.S. degree in electrical engineering from the Ecole Nationale d'Ingénieurs de Sfax (ENIS), Tunisia, in 2005 and 2007, respectively. In September 2008, he is inscribed in doctor's degree in the field of electrical engineering at the Ecole Nationale d'Ingénieurs de Sfax (ENIS), Sfax, Tunisia. Since 2007, he is an Assistant teacher at the Institut Supérieur d'Informatique de Médenine, Tunisia. His current research interests include electric machines, fault detection and localization, observation and sonsorless control.

Kamel Jarray was born in Ben Guerdane, Tunisia, on June 18, 1967. He received the B.S. and DEA degrees from the Ecole Supérieure des Sciences et Techniques de Tunis (ESSTT), Tunis, Tunisia, in 1993 and 1995, respectively, and the Ph.D. degree from Aix-Marseille III University, Marseille, France, in 2000, all in electrical engineering. From 2000 to 2001, he was a Researcher at the Ecole Supérieure d'Ingénieurs de Marseille (ESIM). From 2001 to 2003, he was an Assistant Professor at the Ecole Supérieure des Sciences et Techniques de Tunis (ESSTT), Tunisia. Since September 2003, he has been an Assistant Professor at the Ecole Nationale d'Ingénieurs de Gabès (ENIG), Gabès, Tunisia. His research is in the areas of electrical machines, sensorless vector control ac motor drives, and advanced digital motion control.

Yassine Koubaa is the Head of Automatic Control Research Laboratory (Sfax-Tunisia) and the Editor in Chief of International Journal on Sciences and Techniques of Automatic control & computer engineering (IJ-STA). He received the B.S.and DEA (master) degrees in 84 and 86, respectively, the Doctorat theses in 1996, the "Habilitation Universitaire" (HDR) from the National Engineering school of Sfax (ENIS) all in Electrical Engineering. From 1989 to 1996, he was an Assistant Professor in the Electrical Engineering Department of ENIS. In September 1997, he became an Associate Professor. Since September 2005, he is a full professor at the same university. His main research interests cover several aspects related to electrical machines, including systems identification, advanced motion control and diagnostics. He has authored more than 70 papers in international conferences and technical Journals in the area as well as many patents. He serves as a member of the Scientific and the Technical Program Committees of several international conferences and technical Journals in the motor drives fields.

Mohamed Boussak currently serves as a Member of the Technical program committees of several international conferences and scientific Journals in the areas of power electronics and motor drives fields. He received the B.S. and DEA degrees from the Ecole Normale Supérieure de l'Enseignement Technique de Tunis (ENSET), Tunisia, in 1983 and 1985 respectively, the Ph.D.degree from Pierre et Marie Curie University (Paris 6), Paris, France, in 1989, the "Habilitation à Diriger des Recherches" (HDR) from Aix-Marseille III University, Marseille, France in 2004, all in Electrical Engineering. From September 1989 to September 1990, he was a Researcher with the Ecole Supérieure d'Ingénieurs de Marseille ESIM. From October 1990 to September 1991, he was a Research Teacher in Electrical Engineering with the Claude Bernard University, Lyon, France. From October 1991 to June 2004, he was an Associate Professor with the Ecole Supérieure d'Ingénieurs de Marseille (ESIM), France. From July 2004 to December 2008, he was an Associate Professor of Electrical Machines with the Ecole Centrale Marseille (ECM), France, where, since January 2009, he has been a Senior Professor. His research areas, in the Laboratoire des Sciences de l'Information et des Systèmes (LSIS), UMR CNRS 6168, Marseille, France, are electrical machines, power conversion systems, sensorless vector control ac motor drives, advanced digital motion control and diagnostic for industrial electric system. He has published more than 100 papers in scientific Journals and conference proceedings in these research fields. Dr. Boussak is Senior Member of IEEE Industry Application, IEEE Industrial Electronics and IEEE Power Electronics Societies.

FHESMM: Fuzzy Hybrid Expert System for Marketing Mix Model

Mehdi Neshat[1], Ahmad Baghi[2] ,Ali Akbar Pourahmad[3],Ghodrat Sepidnam[4] ,Mehdi Sargolzaei[5],Azra Masoumi[6]

[1] Department of Computer, Shirvan Branch, Islamic Azad University, Shirvan, Iran

[2] Department of management science , Shirvan Branch, Islamic Azad University, Shirvan, Iran

[3] Department of information and library science, Shirvan Branch, Islamic Azad University, Shirvan, Iran

[4]Department of Computer, Shirvan Branch, Islamic Azad University, Shirvan, Iran

[5] Department of Computer, Shirvan Branch, Islamic Azad University, Shirvan, Iran

[6]Department of Computer, Shirvan Branch, Islamic Azad University, Shirvan, Iran

Abstract

Increasing customers' satisfaction in this developed world is the most important factor to have a successful trade and production. New marketing methods and supervising the marketing choices will have a key role to increase the profit of a company. This paper investigates an expert system through four main principles of marketing (price, product, Place and Promotion) and their composition with a logic fuzzy system and benefiting from the experiences of marketing specialists. Comparing with the other systems, this one has special properties such as investigating and extracting different fields in which affect the customers' satisfaction directly or indirectly as input parameters (26), using knowledge of experts to design inference system rule, composing the results of five fuzzy expert systems and calculating final result(customer's satisfaction) and finally creating a high function expert system on management and guiding the managers to do a successful marketing in dynamic markets.

Keywords: *marketing mix model, fuzzy decision making system, fuzzy, expert system, mamdani inference, four P's.*

1. Introduction

Philip Kotler has defined marketing management as the analysis, planning, implementation, and control of programs designed to bring about desired exchanges with target audiences for the purpose of personal or mutual gain [1]. One of the most critical marketing management decisions is that decision of setting the marketing mix values, and selecting and employing strategy that periodically changes that marketing mixes in response to changing business environment. The marketing mix problem involves setting the values of the marketing decision variables; the four P's; namely, Product, Price, Place and Promotion. Developing an effective marketing mix is important for product planners seeking to gain competitive advantage in industrial markets. The decision regarding specifying the marketing mix depends on a set of variables, the majority of which are **stochastic**, **dynamic**, **vague** or **inexact**, and **qualitative** or **intangible;** such as competitor's price, competitor's product quality, competition level, forecasted sales and others**.** These types of variables necessitate adoption of appropriate approaches that can deal with such variables' nature. These variables natures are inherent in various business sectors, specially in case of agriculture business, like agro-food companies, producers of fertilizers, and other agro-chemical products, where the existence of some stochastic variables such as climate, forecasts, demand and a varieties of qualitative variables like food safety, availability, competition, etc. The proposed model is generally applicable to any business sector or industry and specially useful and appropriate in the situation where stochastic, qualitative and vague variables are inherent in the inputs to the problem [2].

2. Literature Review

Traditionally, the problem of setting the marketing- mix has been dealt with in a partial manner, in the sense that most of the articles considered only one element of such mix at a time. For instance, in 1987 Magruth and Kenneth provided three major criteria for evaluating marketing channels [3]. In 1989, Lyrch and Hooky explored the question of possible changes in industrial advertising practice by focusing on the advertising budgeting approaches revealed in recent large-scale U.K. survey [4]. In 1995, Earl Cox described a model for new product pricing [5]. Fuzzy logic methods are very important at management and subgroups. Z. L. Yang Use of Fuzzy Evidential Reasoning in Maritime Security Assessment [14] and Enrico Zio a Fuzzy Decision Tree for Fault Classification [15], Enrico Cameron Risk Management and the Precautionary Principle: A Fuzzy Logic Model [16]. The model combines the expertise of financial, marketing, sales, and manufacturing management to develop a recommended initial pricing position for a new consumer product. This pricing model showed how fuzzy rule-based system can combine the intelligence of several experts into a single, cohesive process. Little literature attempted to deal with the stochastic, vague and qualitative nature of variables, which inherently affects such marketing decision or provide a whole method for setting the four P's and also very little ones that have considered the practical expression of product quality and integrating it with other 3 P's. However, in attempting to treat the problem from a total perspective, Bay Arinze in 1990 described a computer-based marketing decision support system to support planning strategy for marketing and as an expert system shell aid in the selection of marketing mix variables' values [6]. In 1992, Arinze and Burton developed a simulation model as the heart of a marketing decision support system (MKDSS) to model the stochastic element of the marketing mix, marketing dynamics, the interactions between marketing instruments and competitive effects, to support decision making process and developing the marketing mix [7]. In 2001, Fazlollani and Vahidov attempted to extend the effectiveness of simulation-based DSS through genetic algorithms [8]. They applied a hybrid method based on the combination of Mont Carlo Simulation and GA to the marketing mix problem to improve the process for searching and evaluating alternatives for decision support. Genetic algorithms for tourism marketing was proposed by stephen hurley[8]and Bayesian neural network learning for repeat purchase modeling in direct marketing was proposed by bart baesens[9].modeling fuzzy data in qualitative marketing research was proposed by sajeed varki[10] On optimal partially integrated production and marketing policy with variable demand under flexibility and reliability considerations via Genetic Algorithm [11]and Variable selection in clustering for marketing segmentation using genetic algorithms[12].
2

3. Method

3.1 Fuzzy Marketing Method

Marketing management and Customer Relationship Management (CRM) need methods to analyze, evaluate and segment their customers according their value for the company in order to improve customer relationships, optimize customer or marketing performance and to maximize profitability. One problem of scoring methods like the RFM (Recency, Frequency, Monetary value) model, ABC and portfolio analysis is that they have always been applied in a sharp manner so far, i.e. values are assigned sharply to predefined classes. This often leads to misclassifications (under-/overvaluations) [13].

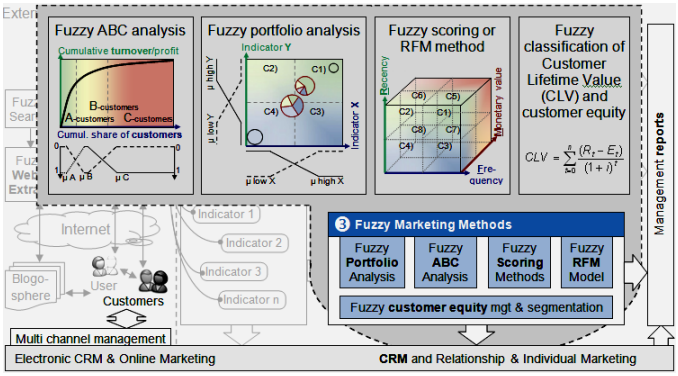


Fig. 1 Fuzzy Marketing Model.

With the fuzzy classification approach, these problems can be avoided, objects are classified exactly and resources can be allocated optimally. In a fuzzy ABC analysis, customers can partly belong to two classes, in fuzzy portfolio analysis to four and in fuzzy scoring methods to several classes at the same time. In addition, the membership degree to each class can be computed, which allows e.g. the calculation of individual, personal prices, accounts or incentives and the adoption of the marketing mix (mass customization).In addition, the fuzzy logic approach can be used also in the domain of performance measurement in order to analyze, classify, evaluate and manage different marketing relevant measures and indicators, for instance customer equity or Customer Lifetime Value (CLV) The main advantage of a fuzzy classification compared to a classical one is that an element is not limited to a single class but can be assigned to several classes. In addition, fuzzy classification and fuzzy methods support qualitative and quantitative indicators. The fuzzy

classification with its query facility allows improving customer equity, launching loyalty programs, automating mass customization issues, and refining marketing campaigns.

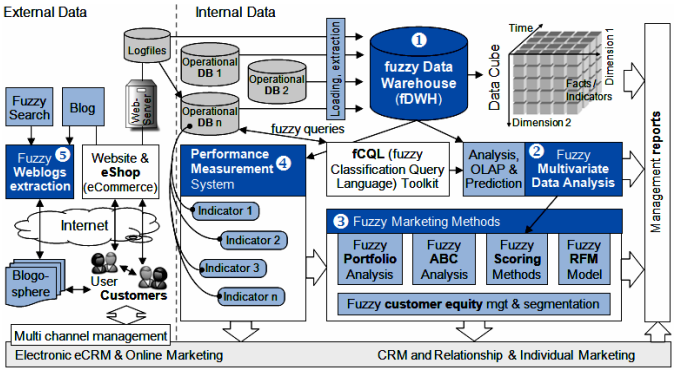


Fig. 2 The framework of fuzzy marketing methods.

3.2 The aim and new method

The marketing-mix problem is a typical problem, which involves vague and uncertain type of input variables and dynamic, non-linear relationships. The problem involves setting the values of the marketing decision variables; the four P's (Product, Price, Place - distribution expenditures and Promotion.

A fuzzy marketing mix model is used in this article this system includes four subsystems having several inputs each. The outputs of these systems are considered as input of the fifth system and final result shows the successfulness level of marketing. Other inputs and their characteristics are as follows:

Table 1: characteristics and input of place expert system

Name='marketing place'	Inputs
Type='mamdani'	1)'Export.drop.shippers'
NumInputs=6 NumOutputs=1	2)'Export.merchants'
NumRules=15	3)'ETC'
AndMethod='min' ImpMethod='min'	4)'manufacturer.export.agent'
OrMethod='max'	5)'EMC'
AggMethod='max'	6)'export.brokers'
DefuzzMethod='centroid'	

Table 2: characteristics and input of price expert system

Name='marketing price'	Inputs
Type='mamdani'	1)'request.inducement'
NumInputs=7 NumOutputs=1	2)'price.importance'
NumRules=25	3)'price.quality'
AndMethod='min' ImpMethod='min'	4)'price.adversary'
OrMethod='max'	5)'price.specific.clientele'
AggMethod='max'	6)'price.cast'
DefuzzMethod='centroid'	7)'price.without'

Table 3: characteristics and input of product expert system units.

Name='marketing product'	Inputs
Type='mamdani'	1)'quality'
NumInputs=6 NumOutputs=1	2)'features'
NumRules=20	3)'packaging'
AndMethod='min' ImpMethod='min'	4)'design'
OrMethod='max'	5)'aftersale.service'
AggMethod='max'	6)'lifetime.warranty'
DefuzzMethod='centroid'	

Table 4: characteristics and input of promotion expert system units.

Name='marketing promotion'	Inputs
Type='mamdani'	1)'personal.sale'
NumInputs=7 NumOutputs=1	2)'pictorial.sale'
NumRules=30	3)'radio.sale'
AndMethod='min' ImpMethod='min'	4)'newspaper.sale'
OrMethod='max'	5)'poster.sale'
AggMethod='max'	6)'caption.poster.sale'
DefuzzMethod='centroid'	7)'award.sale'

The relationship among targets, economic conditions, developments, and other input variables from one side and the marketing-mix setting in the other side is non-linear and difficult or cannot exactly defined unless it is expressed in forms of experts' If-Then decision rules. It is now clear and evident that one way to handle all such aspects of the marketing mix problem is the use of fuzzy logic sets, which effectively handle such vague, uncertain, subjective inputs and efficiently model nonlinear relationships between problem inputs and outputs.

3.2.1 Hybrid fuzzy expert systems designing

A fuzzy expert system is a mix of expert and logic fuzzy systems. This system includes five main sections.

1. Expert: who have specialty in a field. This specialization could be experienced or be gained through wide studies.

2. Fuzzification: The fuzzification comprises the process of transforming crisp values into grades of membership for linguistic terms of fuzzy sets. The membership function is used to associate a grade to each linguistic term... It has different types in which the triangle and trapezoid one are used.

3. Inference engine: it draws a conclusion from data and rules based on a field mamdani inference system.

4. Fuzzy rules bank: it is a complex of < If … Then> rules.

5. Defuzzification: the process of producing a quantifiable result in fuzzy logic. Typically, a fuzzy system will have a number of rules that transform a number of variables into a "fuzzy" result, that is, the result is described in terms of membership in fuzzy sets. For example, rules designed to decide how much pressure to apply might result in "Decrease Pressure (15%), Maintain Pressure (34%), and Increase Pressure (72%)". Defuzzification would transform this result into a single number indicating the change in pressure. The simplest but least useful defuzzification method is to choose the set with the highest membership, in this case, "Increase Pressure" since it has a 72% membership, and ignore the others, and convert this 72% to some number. The problem with this approach is that it loses information. The rules that called for decreasing or maintaining pressure might as well have not been there in this case. A useful defuzzification technique must first add the results of the rules together in some way. The most typical fuzzy set membership function has the graph of a triangle. Now, if this triangle were to be cut in a straight horizontal line somewhere between the top and the bottom, and the top portion were to be removed, the remaining portion forms a trapezoid. The first step of defuzzification typically "chops off" parts of the graphs to form trapezoids (or other shapes if the initial shapes were not triangles). For example, if the output has "Decrease Pressure (15%)", then this triangle will be cut 15% the way up from the bottom. In the most common technique, all of these trapezoids are then superimposed one upon another, forming a single geometric shape. Then, the centroid of this shape, called the fuzzy centroid, is calculated. The x coordinate of the centroid is the defuzzified value.

6. Knowledge engineering: KE is an engineering discipline that involves integrating knowledge into computer systems in order to solve complex problems normally requiring a high level of expertise. At present, it refers to the building, maintaining and development of knowledge-based systems. It has a great deal in common with software engineering, and is used in many computer science domains such as artificial intelligence including databases, data mining, expert systems, decision support systems and geographic information systems. Knowledge engineering is also related to mathematical logic, as well as strongly involved in cognitive science and socio-cognitive engineering where the knowledge is produced by socio-cognitive aggregates (mainly humans) and is structured according to our understanding of how human reasoning and logic works.

- Various activities of KE specific for the development of a knowledge-based system:
  - Assessment of the problem
  - Development of a knowledge-based system shell/structure
  - Acquisition and structuring of the related information, knowledge and specific preferences (IPK model)
  - Implementation of the structured knowledge into knowledge bases
  - Testing and validation of the inserted knowledge
  - Integration and maintenance of the system
  - Revision and evaluation of the system.

Being still more art than engineering, KE is not as neat as the above list in practice. The phases overlap, the process might be iterative, and many challenges could appear. Recently, emerges meta-knowledge engineering as a new formal systemic approach to the development of a unified knowledge and intelligence theory.

7. Expert (marketing expert): An expert, more generally, is a person with extensive knowledge or ability based on research, experience, or occupation and in a particular area of study. Experts are called in for advice on their respective subject, but they do not always agree on the particulars of a field of study. An expert can be, by virtue of credential, training, education, profession, publication or experience, believed to have special knowledge of a subject beyond that of the average person, sufficient that others may officially (and legally) rely upon the individual's opinion. Historically, an expert was referred to as a sage.

8. Market or the environment that the expert acts inside it.

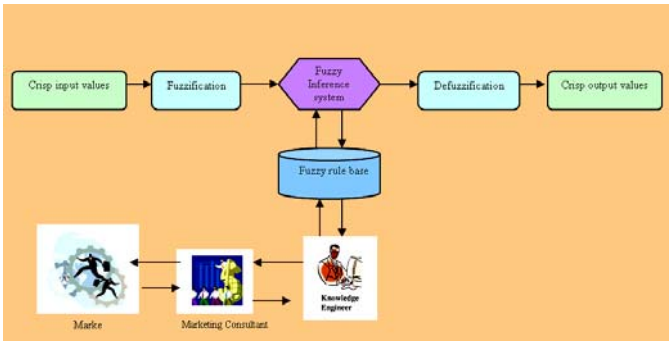


Fig. 3 shows a fuzzy expert system for marketing

This research uses five fuzzy expert systems. Figure 4 shows the system including a fuzzy expert system for each p and the whole results on another system and finally the rate of successfulness.

3.2.1.1 Defuzzification

The values of input and output variables are fuzzified Based on opinion of experts and analysts, triangular

membership functions with five fuzzy sets are used. Except for the variable competition level, five fuzzy sets are used for all other variables: “Very Low” as VL, “Low” as L, “Medium” as M, “High” as H, and “Very High” as VH. The membership functions of all the marketing place fields have been defined below:

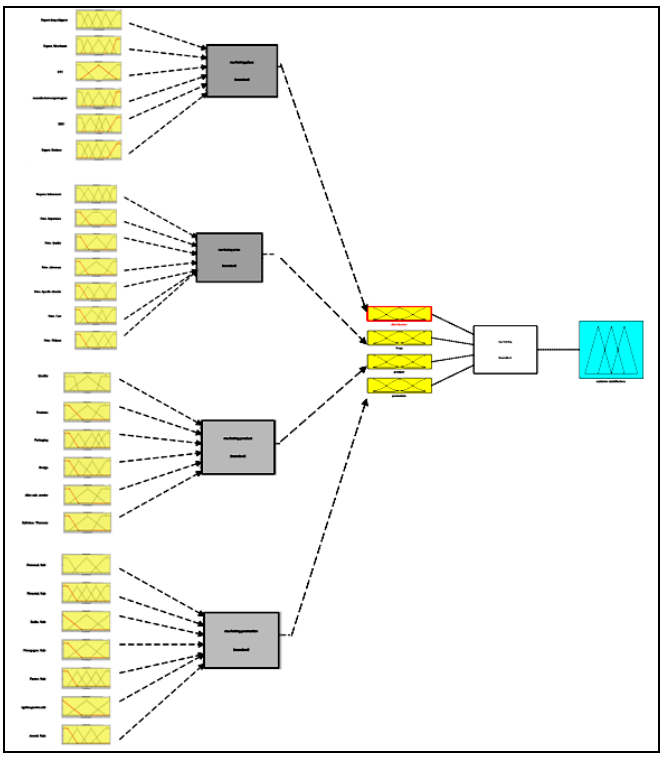


Fig 4.fuzzy hybrid expert system for marketing

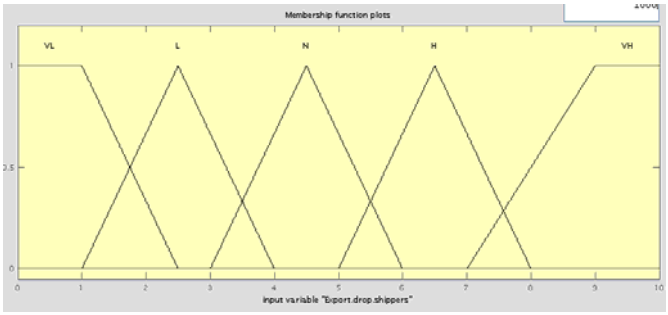


Fig 5.Fuzzification of 'Export.drop.shippers'

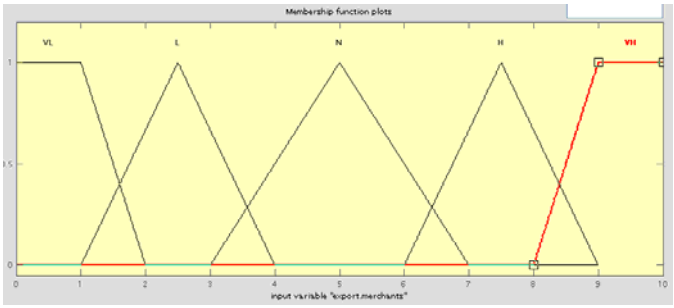


Fig 6.Fuzzification of export. Merchants

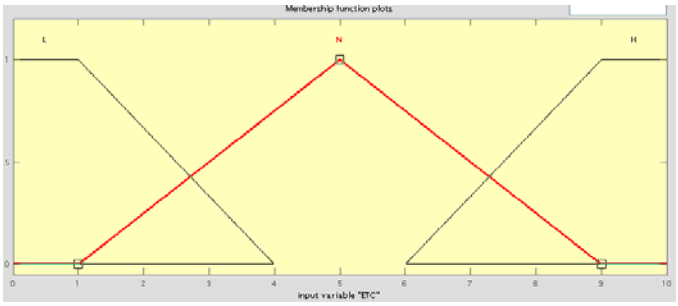


Fig 7.Fuzzification of ETC

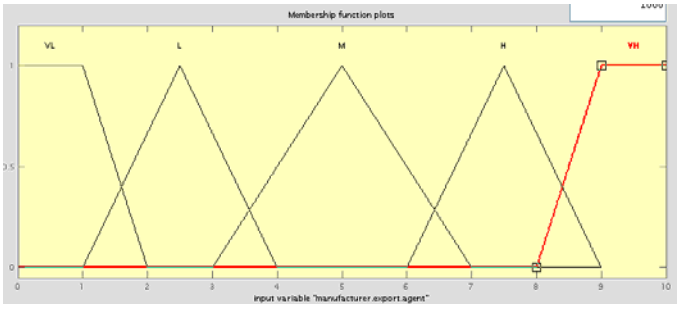


Fig 8.Fuzzification of manufacturer.export.agent

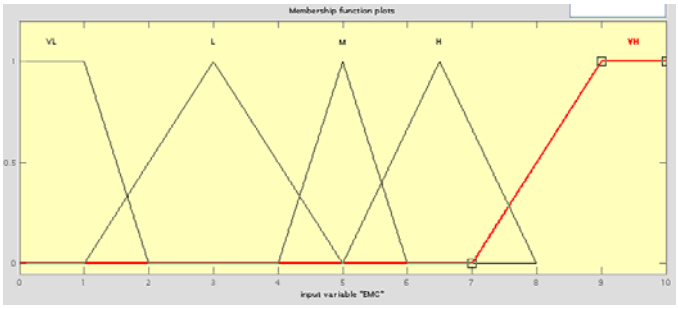


Fig 9.Fuzzification of EMC

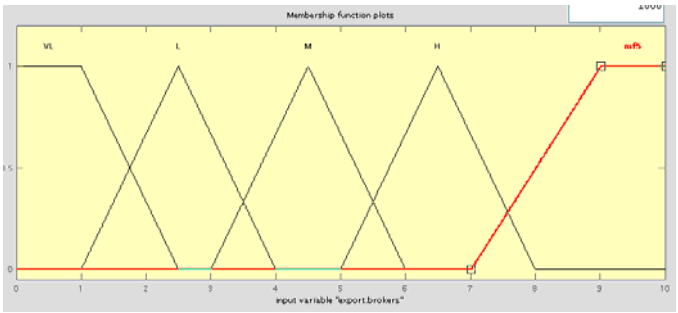


Fig 10.Fuzzification of export brokers

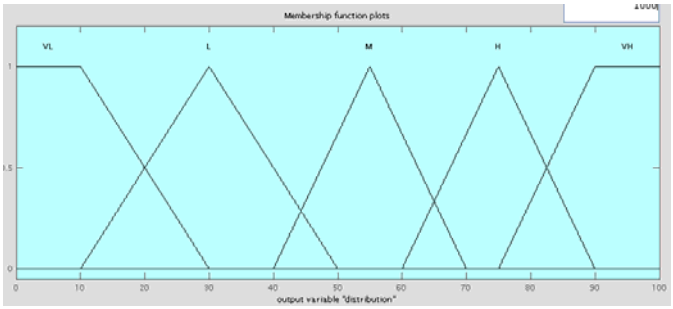


Fig 11.Fuzzification of output distribution

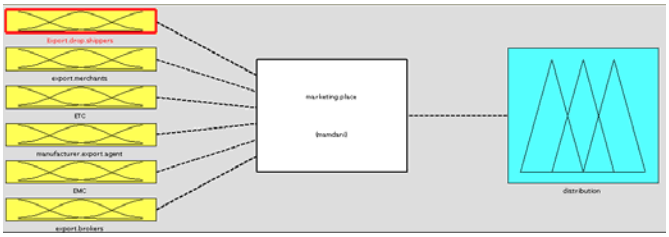


Fig 12.Outline model of place fuzzy expert system

There are several fuzzy statuses such as Very Low (VL), Low (L), Normal (N), High (H) and Very High (VH) for output of place fuzzy expert system that modifies the quality of the product. Regarding triangle fuzzification, High, Low and Very high status are as follows:

$$\mu_{low}(distribution)=\begin{cases}0 & \alpha\le 10\\ (\alpha-10)/20 & 10<\alpha\le 30\\ (50-\alpha)/20 & 30<\alpha\le 50\\ 0 & \alpha>50\end{cases} \quad (1$$

$$\mu_{high}(distribution)=\begin{cases}0 & \alpha\le 60\\ (\alpha-60)/15 & 60<\alpha\le 75\\ (90-\alpha)/15 & 75<\alpha\le 90\\ 0 & \alpha>90\end{cases} \quad (2$$

$$\mu_{veryhigh}(distribution)=\begin{cases}0 & \alpha\le 75\\ (\alpha-75)/15 & 75<\alpha\le 90\\ 1 & 90<\alpha\le 100\\ 0 & \alpha>100\end{cases} \quad (3$$

For example, “Expert.dro.shippers” fuzzy membership function is as follows:

$$\mu_{high}(export.drop.shipper)=\begin{cases}0 & \alpha\le 1\\ (\alpha-1)/4 & 1<\alpha\le 5\\ (9-\alpha)/4 & 5<\alpha\le 9\\ 0 & \alpha>9\end{cases} \quad (4$$

$$\mu_{high(\text{Export.drop.shippers})}=\{\frac{0}{1}+\frac{0.25}{2}+\frac{0.5}{3}+\frac{0.75}{4}+\frac{1}{5}+\frac{0.75}{6}+\frac{0.5}{7}+\frac{0.25}{8}+\frac{0}{9}\} \quad (5$$

3.2.1.2 Rules Bank and Defuzzification

Taking into account different conditions of CS and even situations that have not yet occurred but may occur in the future, the rules have been edited. In total, there are 105 dependent rules, where each rule is a collection of variants that have occurred “AND” together and show a special situation of CS. These rules cover all the situations that the fuzzy system may face. Also, there may occasionally be an opposition be between the base rules. This problem is solved by the inference engine and defuzzification parts of the system. The Inference engine and defuzzification parts give us an optimized result by taking an average of the attained rules. Defuzzification’s centre of gravity formula is used for calculating the certain output amount.

$$D^{*}=\frac{\int D.\mu_{middle}(D)dD}{\int \mu_{middle}(D)dD} \quad (6$$

Table 5.Collection rules of marketing place fuzzy expert system

Export.drop .shippers	Export. merchants	ETC	Manufacturer. Export. agent	EMC	Export. brokers	distribution
(VL)	L	VL	VL	L	VL	VL
VL	L	VL	VL	VL	L	VL
N	N	L	N	N	VL	N
H	N	H	H	H	N	H
….	.....	.....	.....	.....	.....	.....
VH	VH	N	VH	VH	H	VH

Table 6.COLLECTION RULES OF 'marketing price' FUZZY EXPERT SYSTEM

Request .inducement	Price .importance	Price. quality	Price. adversary	specific_ .clientele	Price .cast	Price .without	distribution
VL	VL	VL	L	L	VL	VL	VL
L	L	VL	N	L	VL	L	L
L	N	H	N	N	N	N	N
H	H	H	L	H	H	VH	H
.....	.....	.....	.....	.....	.....	.....	.....
VH	H	VH	VH	VH	H	H	VH

Table 7.COLLECTION RULES OF 'marketing product 'FUZZY EXPERT SYSTEM

quality	features	packaging	design	Aftersale .service	Lifetime .warranty	product
N	VL	VL	VL	L	VL	VL
VL	N	VL	VL	L	VL	VL
L	L	N	L	L	VL	L
N	L	H	N	N	N	N
VH	VL	H	H	H	H	H
.....	.....	.....	.....	.....	.....	.....
H	VH	VH	H	VH	VH	VH

Table 8.COLLECTION RULES OF 'marketing promotion' FUZZY EXPERT SYSTEM

Personal .sale	Pictorial .sale	Radio .sale	Newspaper .sale	Poster .sale	Caption .poster.sale	Award .sale	promotion
VL	VL	VL	N	VL	L	VL	VL
L	L	VL	N	L	L	L	L
N	N	N	N	VL	H	N	N
H	H	H	VH	N	H	H	H
.....	.....	.....	.....	.....	.....	.....	.....
VH	VH	VH	H	VH	VH	H	VH

The accuracy of α rules should be clarified at this stage. Firstly, the minimum amount of each rule is recognized and then the maximum amount between them is chosen. For instance (Export.drop.shippers =2.2, export. merchants =6.6, ETC =3.5, manufacturer.export.agent =3.5, EMC =5.2. export. brokers =6) make rules 22 and 23 active.

$$\alpha_{22} = \min(VL, N, L, L, M, M)$$
$$\alpha_{22} = \min(0.2, 0.2, 0.17, 0.34, 0.8, 0) = 0.17$$

$$\alpha_{23} = \min(L, H, N, N, H, H)$$
$$\alpha_{23} = \min(0.8, 0.4, 0.625, 0.25, 0.13, 0.66) = 0.13$$

Using the mamdani inference (max, min), the system's membership function is:

$$\max(\alpha_{22}, \alpha_{23}) = \max(0.17, 0.13) = 0.17$$

Inference rules for the variables and the output are as follows:

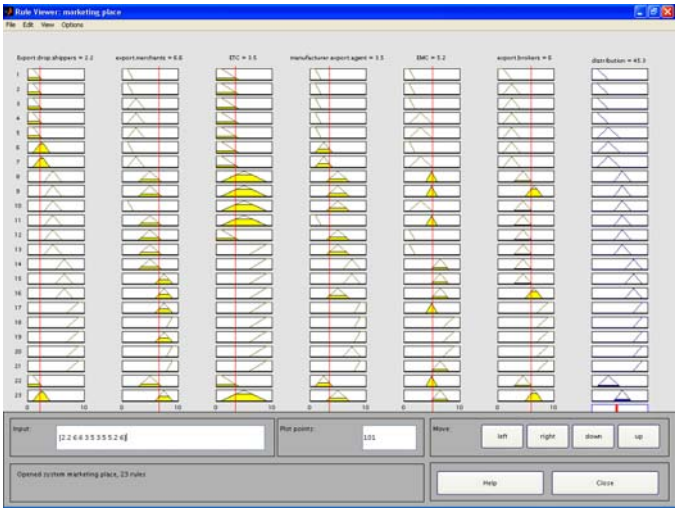


Fig 13.Rule viewer marketing place

For instance (Request. inducement =2.8, Price .importance =7.5, price. quality =5, price. adversary =7.2, specific. clientele =5, Price .cast =6.5, Price .without=1) make rules 6 and 28 active.

$$\alpha_6 = \min(VL, M, N, N, L, M, L)$$
$$\alpha_6 = \min(0.1, 0.5, 1, 0.45, 0, 0.167, 1) = 0.1$$
$$\alpha_{28} = \min(L, H, N, H, M, H, L)$$
$$\alpha_{28} = \min(0.9, 0.4, 1, 0.1, 0.5, 1, 1) = 0.1$$
$$\max(\alpha_6, \alpha_{28}) = (0.1, 0.1) = 0.1$$

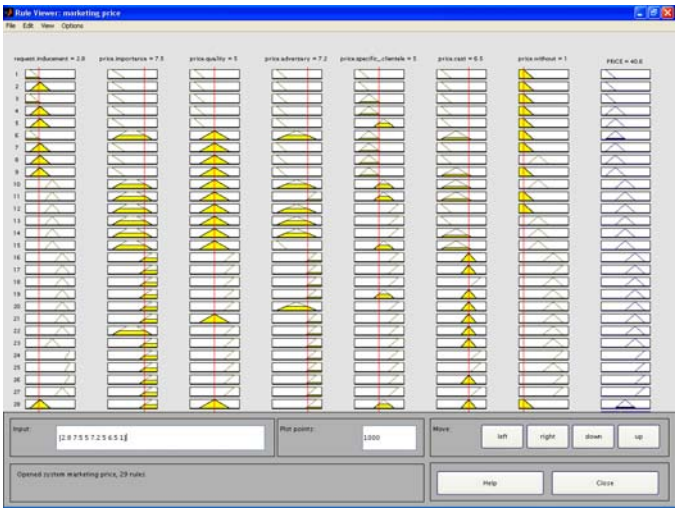


Fig 14. Rule viewer marketing price

For instance (quality =4, features =8.1, packaging =6.3, design =7.5, aftersale.service =9, Lifetime. warranty =7.5) make rules 56 and 71 active.

$$\alpha_{56} = \min(N, H, N, H, N, M)$$
$$\alpha_{56} = \min(0.75, 0.225, 0.85, 0.5, 0, 0.375) = 0$$
$$\alpha_{71} = \min(N, VH, H, VH, H, H)$$
$$\alpha_{71} = \min(0.75, 1, 0.13, 0.25, 1, 0.75) = 0.13$$
$$\max(\alpha_{56}, \alpha_{71}) = \max(0, 0.13) = 0.13$$

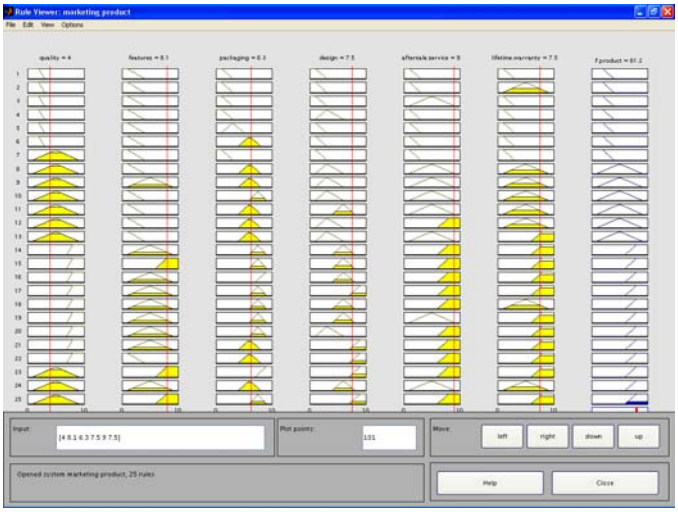


Fig 15.Rule viewer marketing product

For instance (personal. sale=7, pictorial. sale =6, radio. sale=5, newspaper. sale=4, poster. sale=6, caption.poster.sale=7, award. sale=4.5) make rules 56 and 71 active.

$$\alpha_{74} = \min(N, M, N, L, N, N, N)$$
$$\alpha_{74} = \min(0.33, 0.5, 1, 0, 0.5, 0.5, 0.25) = 0.25$$
$$\alpha_{98} = \min(H, H, N, N, H, H, H)$$
$$\alpha_{98} = \min(0.33, 0.5, 1, 0.75, 0.5, 0.25, 0.25) = 0.25$$
$$\max(\alpha_{74}, \alpha_{98}) = \max(0.25, 0.25) = 0.25$$

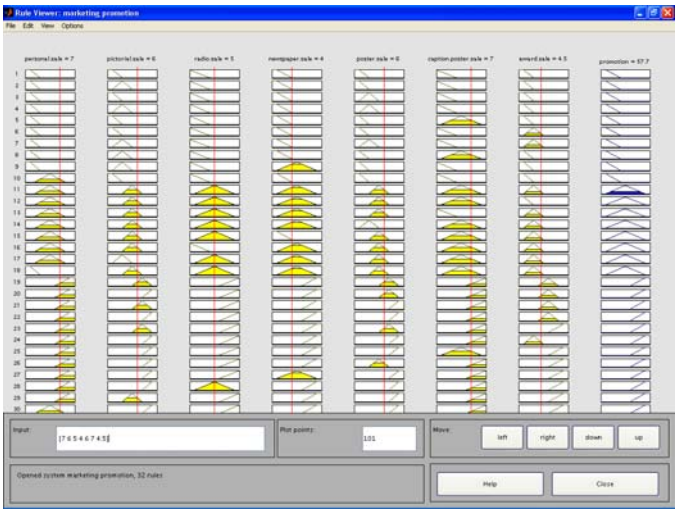


Fig 16.Rule viewer marketing promotion

The fields in final fuzzy expert system are as follows:

For instance (distribution=45.3, price=40.6, product=81.2, promotion=57.7) make rules 6,7,8,14,22 active.

$$\alpha_6 = \min(M, N, M, N)$$
$$\alpha_6 = \min(0.354, 0.53, 0.22, 0.81) = 0.22$$
$$\alpha_7 = \min(L, N, M, N)$$
$$\alpha_7 = \min(0.235, 0.53, 0.22, 0.81) = 0.22$$
$$\alpha_8 = \min(M, L, M, N)$$
$$\alpha_8 = \min(0.354, 0.47, 0.22, 0.81) = 0.22$$
$$\alpha_{14} = \min(L, L, H, N)$$
$$\alpha_{14} = \min(0.235, 0.47, 1, 0.81) = 0.235$$
$$\alpha_{22} = \min(M, N, H, N)$$
$$\alpha_{22} = \min(0.354, 0.53, 1, 0.81) = 0.354$$
$$\max(\alpha_6, \alpha_7, \alpha_8, \alpha_{14}, \alpha_{22}) =$$
$$\max(0.22, 0.22, 0.22, 0.235, 0.354) = 0.354$$

As figure ( ) shows, the rate of satisfaction is low if fields of fuzzy expert have high variables. In this system it is obvious that price and distribution filed have a very high affect.

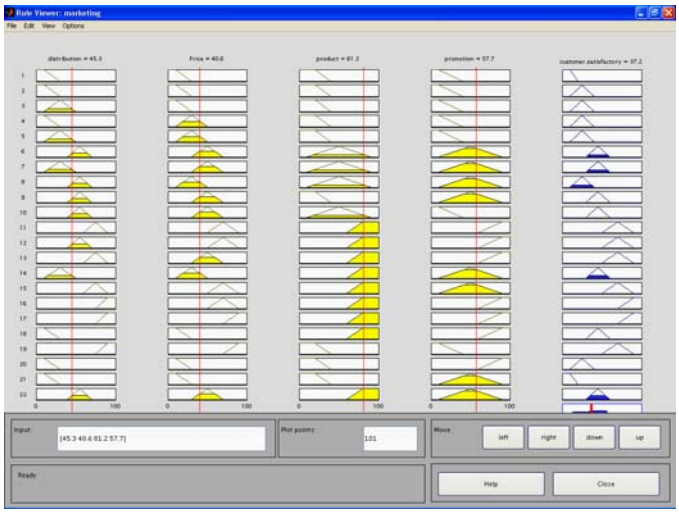


Fig 17.Rule viewer final marketing

4. Conclusions

A mix fuzzy expert system is designed and used to determine the successfulness on marketing based on 4p principle. A fuzzy expert system is designed for each effective field for marketing. It had several inputs and outputs. The results of each four systems input a final fuzzy expert system and show a final logic result through conclusion rules. This result would be considered as an important parameter for experts in marketing that they use it. Any inconvenience on marketing and management will cause several irreparable damages in economy. Therefore, due to marketing risk and its fuzzy nature and meeting the customers having variable behaviors, using this system could be found as a very effective help to prevent the damages. This system applying knowledge and experience of marketing experts could be equipped with very strong inference rules to have very useful and careful results. Comparing the operation of the experts and simulating different situations in the market, this system showed very good result by 91.5% accuracy.

References

[1] Kotler P: Marketing Management: Analysis,Planning, and Control. Prentice-Hall, Inc., Englewood ,1972.

[2] S.Aly,I.Vrana,"fuzzy expert marketing mix model",AGRIC.ECON.-CZECH,51,2005(2),pp 69 - 79.

[3] Magruth A.J., Hardy K.G.: Selecting Sales and Distribution Channels. Industrial Marketing Management Journal, 16 (4):1987, 273− 278, May.

[4] Lyrich J.E., Hooky G.J," Industrial Advertising Budget Approaches in the UK". Industrial Marketing Management Journal,1989, *18* (4): 265–270, November.

[5] Cox E. "Fuzzy Logic for Business and Industry". Charles River Media, Inc, 1995.

[6] Arinze B. " Marketing Planning with Computer Models", A Case Study in the Software Industry.1990, *19* (2): 117–129.

[7] Arinze B., Burton J," A Simulation Model for Industrial Marketing". Omega,1992, *20* (3): 323–335.

[8] Fazolollani B., Vahidov R. (2001): Extending the Effectiveness of Simulation-based DSS through Genetic Algorithms. Journal of Information and Management, *39* (1): 53–65.

[9] B.Baesens,S.Viaene,J.Vanthienen: Bayesian neural network learning for repeat purchase modeling in direct marketing , European journal of operation research 138(2002) 191-121

[10] S.VARKI, B.COOIL, R.T. RUST, modeling fuzzy data in qualitative marketing research, *Journal of Marketing Research* Vol. XXXVII (November 2000), 480–489

[11] P.Pal ,A.K.Bhunia,S.K.Goyal ,: On optimal partially integrated production and marketing policy with variable demand under flexibility and reliability considerations via Genetic Algorithm, applied mathematics and computation 188 (2007)525-537.

[12] H.H.Liu,C.S.Ong: Variable selection in clustering for marketing segmentation using genetic algorithms ,expert system with application 34(2008) 502-510.

[13] http://diuf.unifr.ch/is/fmsquare/

[14] Z. L. Yang , J. Wang , S. Bonsall , and Q. G. Fang: Use of Fuzzy Evidential Reasoning in Maritime Security Assessment,International Journal Risk Analysis , Volume 29, Issue 1, Date: January 2009, Pages: 95-120

[15] Enrico Zio, Piero Baraldi, Irina C. Popescu, A Fuzzy Decision Tree for Fault Classification, International Journal Risk Analysis , Volume 28, Issue 1, Date: February 2008, Pages: 49-67

[16] Enrico Cameron, Gian Francesco Peloso**,** Risk Management and the Precautionary Principle: A Fuzzy Logic Model**,** International Journal Risk Analysis , Volume 25, Issue 4, Date: August 2005, Pages: 901-911.

Mehdi Neshat was born in 1980.He received the B.Sc. degree in computer engineering from Azad University, Maybod, Iran, in 2006, the M.Sc. degree in Artificial Intelligence from the University of mashhad, in 2008 and is a member of the IEEE and the IEEE Computer Society , computer society of Iran , Iranian Fuzzy System Society .

He is with Islamic Azad University, Shirvan Branch, Faculty of Engineering, and computer Engineering Dept., Shirvan /Iran since 2007. His research interests are fuzzy logic, fuzzy systems, and fuzzy neural networks, particle swarm optimization, genetic algorithms, ant colony optimization, and other evolutionary computation techniques. He has publications and submissions in international conferences like applied soft computing, Applied Mathematical Modeling, Expert Systems with Applications, Fuzzy Sets & Systems, Computers in Industry Information Sciences, Mathematical & Computer Modeling.

Ali Akbar Pourahmad was born in 1968.He received the B.Sc. degree in library science from Ahvaz Chamran University, Ahvaz, Iran, in 1991, the M.Sc. degree in information and library science from the Islamic Azad University of Tehran, in 1995 and The PHD Degree in information and library science from the Islamic Azad University, Science and research branch of Tehran in 2004.He's a member of information and library Society of Iran, computer society of Iran. His research interests are digital library, information storage and retrieval, information literacy, indexing and abstracting, information behavior, subject cataloging and organizing information resources and interface of libraries.

Ghodrat Sepidnam was born in 1950. .He received the B.Sc. degree in Electronic Engineering from Ferdowsi University of Mashhad 1970, Iran, in 1974, the M.Sc. degree in electric engineering from the Ferdowsi University of Mashhad, in 1979 the PHD Degree in Electronic Engineering from the Manchester University of Landon, England. He's a member of computer society of Iran, Iranian Fuzzy System Society. He's with Department of computer engineering faculty of engineering.

Mehdi sargolzae was born in 1978.He received the B.Sc. degree in computer engineering from Ferdowsi University, mashhad, Iran, in 2005, the M.Sc. degree in computer engineering from the Amirkabir University of Tehran, in 2007 and Dr student in computer engineering from the Amsterdam University.

Azra Masoumi was born in 1981.He received the B.Sc. degree in computer engineering from Azad University, meybod, Iran, in 2008 and is a member computer society of Iran , Iranian Fuzzy System Society .

Design and characterization of tapered transition and inductive window filter based on Substrate Integrated Waveguide technology (SIW)

Nouri Keltouma[1] , Feham Mohammed[1] and Adnan Saghir[2]

[1]Laboratoire de recherche Systèmes et Technologies de l'Information et de la communication STIC, Faculté des Sciences - Université de Tlemcen Algérie.

[2]Laboratoire de Laplace, INP ENSEEIHT, Toulouse France

Abstract

Microstrip tapered transition and inductive band-pass filter around 5 GHz, using Substrate Integrated Waveguide technology (SIW) are studied in this paper. All the structures are designed with Finite Element Method (FEM) and fabricated on a single substrate of Epoxy FR4 using a standard PCB process. The return loss of the proposed filter is better than 20 dB.

The measured results for all the structures investigated show a good agreement with the simulation results.

Keywords: *Substrate Integrated Waveguide (SIW), Band-pass Filter, Transition, Via-Holes, SIW-Microstrip Technology.*

1. Introduction

In recent years, a new waveguide technology called the substrate integrated waveguide (SIW) has been introduced in many microwave communication systems, such as Wireless Local Area Networks (WLAN). This technology has been successfully used to design microwave and millimeter-wave filters which are widely exploited extensively as a key block in modern communication systems [1]-[4].

The traditional rectangular waveguide technologies are used in various microwave and millimeter-wave communication systems, especially communication satellites, earth stations, and wireless base-stations, due to their high Q values and high power capability. However, they are expensive to fabricate, voluminous and do not integrate with planar structures in electronic systems. Microstrip lines, on the other hand, are easy and not expensive to fabricate, but are not low loss radiation and not shielded.

SIW components take the advantages of low radiation loss, high Q- factor and high power in systems. Additionally, they have a small size compared to the corresponding conventional rectangular waveguide components.

They are constructed by metal filled via-hole arrays in substrate and grounded planes which can be easily interconnected with other elements of the system on a single substrate plat form without tuning, this system can be miniaturised into small package called the system in package SIP which has a small size and a low cost [5]. A schematic view of an integrated waveguide is shown in Fig. 1.

In this paper, the finite element method (FEM) based on a commercial software package "HFSS" has been applied to the analysis of the SIW structures. Firstly we propose C-band SIW microstrip line with tapered transition, and then we focus on the design of SIW inductive window filter around 5GHz. The designs of these structures are fabricated by using a low cost printed circuit board (PCB) technology and measured by means of a Vector network analyser (HP8720C).

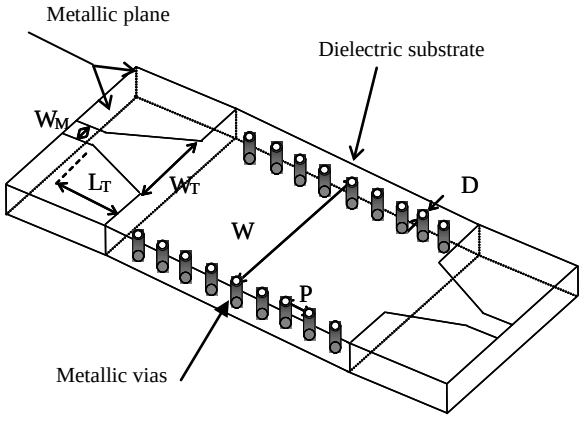
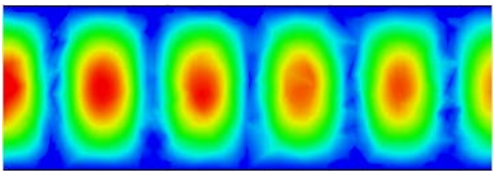


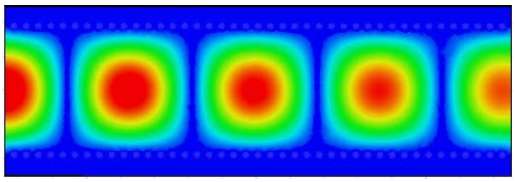
Fig. 1 Topology of the substrate Integrated Waveguide

2. Substrate Integrated Waveguide resonator cavity

A substrate-integrated waveguide (SIW) is made of metallic via arrays in the substrate between top and bottom metal layer replacing the two metal sidewalls.
The propagation properties of the mode in the SIW are very similar to the electromagnetic field distribution of TE_{10}-like mode in a conventional metallic rectangular waveguide (RWG).
In order to compare the electromagnetic field distribution in SIW and rectangular waveguide RWG, we take the diameter of the metallic via D = 1 mm and the period of the vias P = 1.8 mm. The distance between the rows of the centres of via is W = 19 mm. The top, middle, bottom and sidewall metallizations are all copper and the dielectric material is FR4 substrate with $\varepsilon_r = 4.4, \tan\delta = 0.02$.
Fig. 2 shows the cross-sectional view field distribution of dielectric waveguide and SIW without transitions at 5.5 GHz.



a) Rectangular waveguide



b) SIW without transitions

Fig.2 Electric fields distributions in rectangular waveguide and SIW.

We observe that the dominant mode of the SIW resembles the TE_{10} mode of conventional waveguide. The maximum field is present at the middle of the guide.
Thus, the initial dimensions of SIW resonator cavity can be determined by the conventional resonant frequency formula of metallic waveguide resonator, where the length and width of the conventional dielectric waveguide cavity: L_G and W_G, should be replaced by the equivalent width and length of the SIW cavity, L and W, because of the presence of vias sidewall.
So, an SIW cavity can be designed by using the following relations [6]:

$$L = L_G - \frac{D^2}{0.95.P} \quad (1)$$

$$W = W_G - \frac{D^2}{0.95P} \quad (2)$$

In equations (1) and (2), the parameters D and P are the diameter and the period of via holes respectively.

3. A design of proposed transition

In order to combine SIW and microstrip technologies, SIW-microstrip transitions are very required [7]-[8]. Tapered transition shown in Fig. 1 has been studied.
This kind of transition consists of a tapered microstrip line section that connects a 50 microstrip line and the integrated waveguide. The taper is used to transform the quasi-TEM mode of the microstrip line into the TE_{10} mode in the waveguide.
It is known that the propagation constant of the TE_{10} mode is only related to the width "W" .Therefore, the height or the thickness "b" of the waveguide can be reduced without much influence on the TE_{10} mode propagation, thus allowing its integration into a thin substrate that could reduce the radiation loss of the microstrip line.
The design of this transition is very critical and important in order to have a good performance. The optimisation of the transition is obtained by varying the dimension (L_T, W_T) of the tapered geometry (Fig.1).
This structure is fabricated on a substrate FR4 ($\varepsilon_r = 4.4, \tan\delta = 0.02$), the distance between the rows of the centres of via is w = 19 mm, the diameter of the metallic via is D = 1 mm and the period of the vias P = 1.8 mm. The width of tapered W_T is 5.82 mm, its length is L_T = 22 mm. The transitions have been realized using PCB process (fig. 3).

Fig. 3 A Photograph of the manufactured SIW- microstrip line with tapered transitions

This line is simulated by using HFSS, including the two SMA connector's influences. The simulated results and the measured results are compared in Fig. 4.

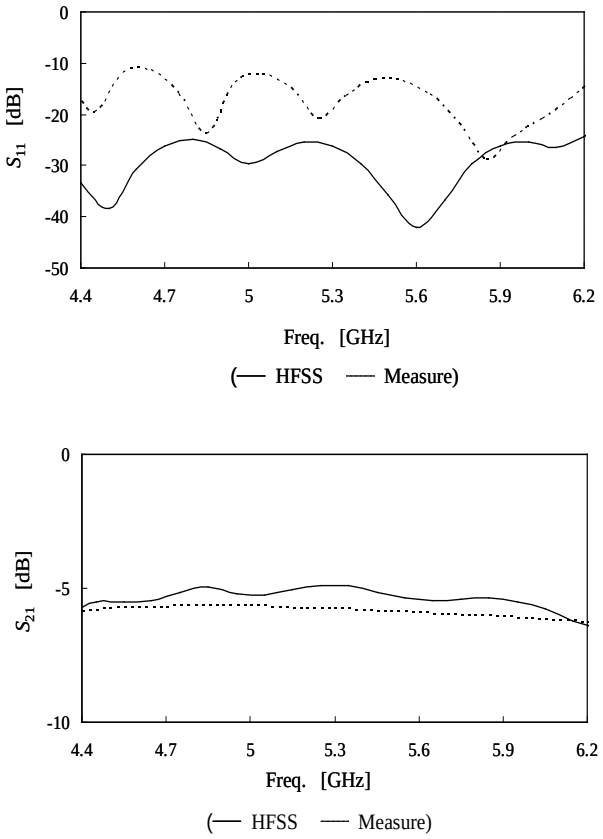


Fig. 4 Return and insertion losses for SIW-microstrip tapered transition

4. A design of SIW filter

4.1 Filter Configuration

Fig.5 Shows the proposed design of filter, this filter includes two microstrip tapered transitions and four SIW resonators cavities.

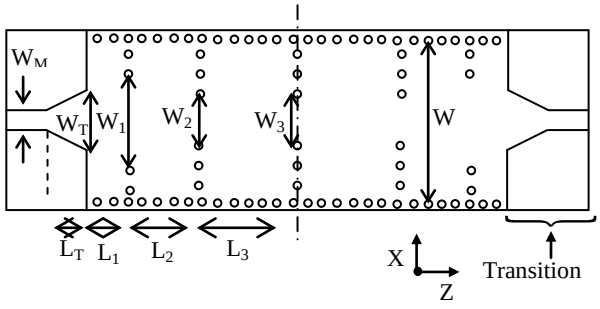


Fig. 5 Geometry of microwave SIW filter

Since the field distribution of mode in SIW has dispersion characteristics similar to the mode of the conventional dielectric waveguide, the design of the proposed SIW band-pass filter, makes use of the same design method for a dielectric waveguide filter. The filter can be designed according to the specifications [9]-[10]. The equivalent circuit of band-pass filter is given by Fig.6.

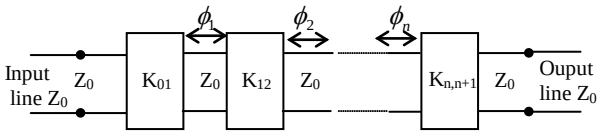


Fig.6 Equivalent circuit of SIW filter

This circuit represents an impedance inverter $K_{n,n+1}$ and a phase shift ϕ_n, the normalized K-inverter values can be calculated as described in [9]-[11] and can be physically realized in terms of discontinuities in a rectangular waveguide using the scattering parameters S_{ij} as follows:

$$K = \sqrt{\frac{1-|S_{11}|}{1+|S_{11}|}} \tag{3}$$

Using this equation, the electrical lengths of the resonators ϕ_i are obtained [11]:

$$\phi_i = \pi - \frac{1}{2}\left(\frac{\pi}{2} - \angle S_{11}\right) \tag{4}$$

$$L_i = \frac{\lambda_g \phi_i}{2\pi} \tag{5}$$

λ_g is the guided wavelength.

4.2 Simulated and experimental results

We designed the proposed SIW filter using the last steps procedure. The initial filter parameters are optimized and given in Table 1.

Table 1: Dimensions of the siw filter

L_i (mm)	5.5	13.6	14.9
W_i (mm)	12.57	10.2	9.6
W_M= 1.4	W_T= 5.8	L_T =21.7	

Each rectangular cavity is created with many rows of via-holes, which have radius of 0.5 mm; the distance between these metallic-vias is set to 1.8 mm. The SIW filter is symmetrical along z axis, it has been fabricated on FR4

substrate ($\varepsilon_r = 4.4, \tan\delta = 0.02$) with a thickness of 0.8 mm.

The filter proposed is done with HFSS using the Finite Element Method (FEM) and fabricated using a standard PCB process (Fig.7) and measured with Vector Network Analyzer. The Fig. 8 shows the comparison between the electromagnetic simulation and the measurements for the input reflection and the transmission, respectively.

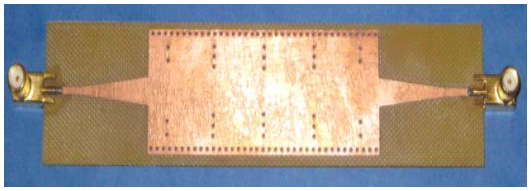


Fig.7 A Photograph of the manufactured SIW filter

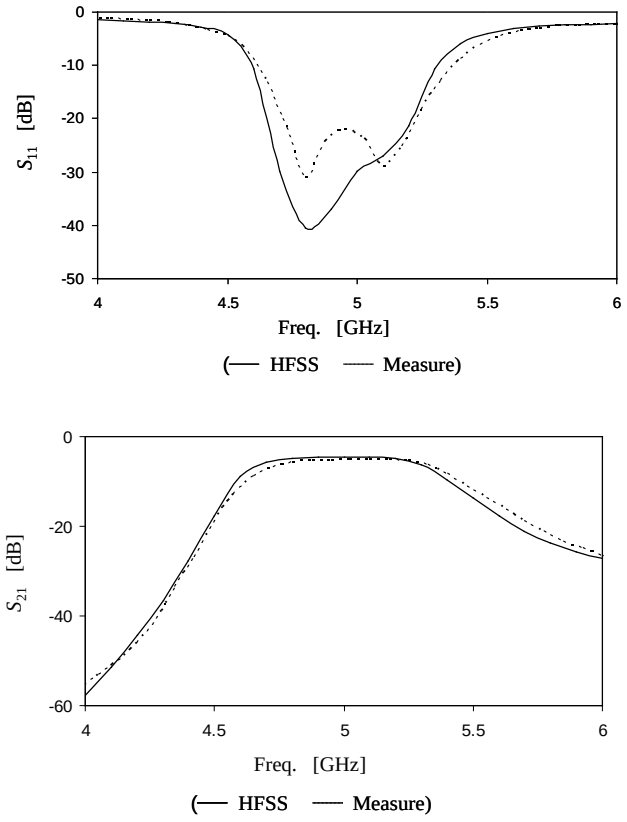


Fig. 8 Return and insertion losses of SIW filter

It can be seen that the agreement is good, the filter has a frequency bandwidth of 750 MHz, the insertion loss around frequency f = 5 GHz is approximately -5 dB the return loss in the pass-band is better than -20 dB.

5. Conclusion

In order to design low cost devices for C band, SIW-microstrip transition and band-pass filter have been treated with HFSS, fabricated and tested. The filter shows a good performance in terms of return and insertion losses. The main characteristics of these kinds of SIW structures are that they have a small size, a high power handling and are easily manufactured.

References

[1] Z. Sotoodeh, B. Biglarbegian,F. H. Kashani and H. Ameri, "A novel bandpass waveguide filter Structure on siw technology," Progress In Electromagnetics Research Letters, Vol. 2, 2008, pp 141–148.

[2] Yujian Cheng, Wei Hong, and Ke Wu, "Half Mode Substrate Integrated Waveguide (HMSIW) Directional Filter," IEEE Microwave and Wireless Components Letters, VOL. 17, July 2007, pp 504-506.

[3] Xiao-Ping Chen and Ke Wu, "Substrate Integrated Waveguide Cross-Coupled Filter with Negative Coupling Structure," IEEE Transactions On Microwave Theory and Techniques, VOL. 56, January 2008, pp 142-149 .

[4] X.-P. Chen, W. Hong, J. Chen, and K.Wu,"Substrate integrated waveguide elliptic filter with high mode," in Proc. Asia–Pacific Microw.Conf., Suzou, China, Dec. 2005.

[5] L. Yan, W. Hong, K. Wu and T.J. Cui, "Investigations on the propagation characteristics of the substrate integrated waveguide based on the method of lines," IEE Proc.-Microw. Antennas Propag, Vol. 152, No. 1, February 2005.

[6] Y. Cassivi, L. Perregrini, P. Arcioni, M. Bressan, K. Wu, and G. onciauro, "Dispersion characteristics of substrate integrated rectangular waveguide," IEEE Microw. Wireless Compon. Lett, Vol. 12, Feb. 2002, pp.333–335.

[7] B.S.Kim, J.W.Lee, K.S.Kim and M.S.Song, "PCB substrate integrated waveguide filter using via fences at millimeter wave", IEEE MTT-S Digest, 2004.

[8] Asanee Suntives, and Ramesh Abhari, "Transition Structures for 3-D Integration of Substrate Integrated Waveguide Interconnects," IEEE Microwave And Wireless Components Letters, VOL. 17, October 2007, pp. 697-699.

[9] Shi Yin, Tatyana Vasilyeva, Protap Pramanick, "Use of three-dimensional in the synthesis of waveguide round rod bandpass filters," International Journal of RF and Microwave CAE, Vol 8, June 1998, pp. 484-497.

[10] Seymour B. Cohn, "Parallel-coupled transmission line resonators filters," IRE Transactions On Microwave Theory and Techniques, April 1958, pp 223-231.

[11] Ji-Fuh Liang and Kawthar A.Zaki, " CAD Micrwave junctions by polynomial curve fitting," IEEE MTT-S Digest, 1993, pp 451-454.

Indirect DNS Covert Channel based on Base 16 Matrix for Stealth Short Message Transfer

M.A. Ngadi, S.N. Omar, and I. Ahmedy

Faculty of Computer Science and Information Systems,
University Technology Malaysia, Skudai, Johor 81310, Malaysia

Abstract

Covert Channel are the methods to conceal a message in the volatile medium carrier such as radio signal and network packets. Until now, covert channels based on the packet length produce abnormal packet length when the length of the message is long. Abnormal packet length, especially in the normal network will expose the covert channels to network security perimeter. Therefore, it motivates the study to propose a new method based on reference matrix to hide the secret message in DNS request. Normal DNS request packet was collected from the campus network. The proposed packets length covert channel was compared with normal DNS request packets. The study found that the new purpose covert channels produce normal DNS packet length according to the campus network.

Keywords: *Covert channels, Packet length, DNS.*

1. Introduction

The encryption is a method where the readable messages are scrambled into unreadable messages through transformations or permutations traditionally or conventionally with a key to protect the information from being readable by unauthorized party [1]. However, encryption alone cannot protect the confidentiality of the message because the unreadable message will attract the attacker to attack the communication channel and try to decrypt the message [2]. Moreover, the encryption itself does not prevent the adversaries from detecting the communication pattern [2].Furthermore, in communication, the encryption itself raises suspicion and triggers further investigation action [3] and knowing there exists a communication between two parties is already valuable information to the attacker [4]. Therefore, these motivate the study to find a method in network communications, where the secret message can be delivered without using an encryption on the network layer up to the application layer and at the same time, preventing the adversary from detecting the communications. It resembles Steganography, where the message is written on a piece of wood and then, waxes the surface of the wood to cover up the message. In network communications, the act of hiding the message in network protocol or communications is called covert channels [5].

Covert Channels are the desirable choices to send secret message based on the stealthiest and volatile of the packets. The stealth is possible to achieve because there are fields in the packet where the characteristics of the fields are random unused or not symmetrically controlled between the network devices within the network where the packets travel; known as Storage Covert Channels [6]. There is also a technique where the data could be hidden between the time arrivals of the packets, known as Timing Covert Channels [6]. The types of Covert Channels depend on the network security perimeter control between the sender and receiver and the stealthiest of the Covert Channels. The later is more preferable because it can resist some security perimeter. Additionally, the volatility of the packet's leverage is the stealthiest against Steganography without leaving any trail to be audited, because the packets will be destroyed after being used or processed according to defined criteria, which further motivates the use of covert channels [5,7]. Moreover, the motivations to use covert channels is supported with the quantity of data that can be transferred through covert channel annually, which can be as huge as 26Gb of data, although the data being transferred is only one bit at a time.8]. Therefore, there is no reason to doubt the capability of the Covert Channels in sending secret messages over the networks.

On the other hand, what makes the covert channels useful is the stealth of the covert channel [9]. The property of the stealth of the covert channels is attributed from the anonymity of covert channels as described in [10], which is subjected to three pillars; plausibility, undetectable and indispensability. Plausibility means the covert channel must be able to exploit the medium in which the packet is in use by the adversary. Undetectable means the amount of the bits sent should not violate the distribution of the normal packet. Indispensability means the adversary must use the medium and will not block the medium on the security perimeter.The indispensability is the most important aspect in stealth property because it will ensure

whether the proposed covert channel is useful or verse versa.

Plausibility and undetectably are co-related. To be truly plausible the covert channels cannot just exploit the packet's unused fields or the randomness of the packets without fulfilling the symmetry of' the packets as in [10,11,12]. In TCP packets, an independent packet could be seen as unfinished TCP handshake activity [13]. In literal meaning, how the data hidden in the packets and the packets operate have a strong correlation with the stealth of the Covert Channel [4,14,15]. Basically, there are two methods to hide and retrieve the data with Storage Covert Channels; indirect and direct methods. Indirect technique hides and retrieves the data by substituting the symbols with the property of the packets whereas the direct technique directly embeds the data into the field of the packet [16]. However, there are ambiguities in the definition of Indirect hidden method.

Hitherto, the direct covert channels have been the favorites and target research in the covert channel's research because the protocol fields to be manipulated are tangible within defined characteristic, based on the survey in [17].

Conversely, direct embedding is not always covert because little attention has been addressed to the communication in the context of the internet [9]. The effect of embedding the data directly into the innocuous protocol fields is it does not hide the fields from traffic analysis, which could map the connection structure to retrieve the uncovered information [9]. Moreover, direct embedding of the data in protocol fields will not be able to resist the network security approach known as the protocol scrubber. Protocol scrubber is a method of an active interposition mechanism to homogenize the network flows by identifying and removing the malicious content in the traffic flow through normalizing the protocol headers, padding and extensions [18]. Therefore, with the sophisticated protocol scrubber mechanism, the direct hidden method could satisfy plausibility of the covertness. However, it could not satisfy the indispensability of covert channels. This, further, forces and motivates the study to look into the indirect hidden method.

Looking back at the previous hidden method, the study is keen on indirect hidden method based on the packet lengths because the packet is directly under the control of the sender. Therefore, there are no malicious attempts to control other state property, and it is not prone to protocol scrubber. However, there are a couple of problems with previous packet lengths. First, the packet length is hard to implement because it could only work in a controlled environment with end-to-end connections as the lengths of the packets depend on the MTU of the routers in the packet travel across the path and the size of MSS. Secondly, as highlighted by Liping in [15], not only the MTU and control environment are the obstacles to packet lengths, but also the normal lengths distribution of the packet lengths.

At this juncture, indeed, the normal distribution of packet lengths is directly associated with the plausibility and undetectably of the covert channels. The plausibility of the packet length's covert channels could not only be achieved when the length of the packets is associated with a value starting from 1 to 256, because, as shown by Liping in [15], the length of the packets is not randomly distributed between 1 to 256. Moreover, there is a gap between the lengths in normal distribution, which is too odd for 256 lengths differently or randomly. Figure 1 is taken from [15], which shows the normal packet lengths distribution of about 2000 packets. The vertical line is the length. In fact, it is clear that, there are only about 20-30 different lengths among the 2000 packets, which will be appropriate for a covert channel with 256 different length distributions as in [19,20,21]. Therefore, the plausibility of the exploited lengths is not convincing in [19, 20, 21] when compared to the normal packet length distribution.

Certainly, if the plausibility of the exploited lengths is not fulfill, the undetectably could not be satisfied. Liping in [15], proposed a method which used a reference of the lengths to overcome the trouble with too many packet length's distributions as in [19,20,21]., The Liping's model used sixteen different length based on the baseline length of the reference agreeable between the sender and receiver. This means, for every 4 bits of the data, there is additional of at least between one to sixteen bytes on the normal packet length. The problems become worse when the sender and receiver have to update the reference length to contain long messages. It would be noticeable that when the size of the message is huge, the Liping's method turns out from normal length distribution to abnormal length distribution as mentioned in [22]. Therefore, the Liping's method cannot satisfy the plausibility of the distribution of the packet length when the message is too long.

This is another challenge that this research would like to address. With the above problem, this research proposed a method, that whereby, there is no additional length of packets that needs to be added to deliver the message as in Liping's method, and the association of the packet lengths is not subject to one to one association as in [19,20,21] methods. The proposed method will allow an association of one to many. That is , the proposed method will introduce a reference, instead to the length of packet. It

will also, take the data in the packet payload as a reference. With this, it will not produce an odd packet length's distribution and there is no additional packet length to be added, therefore, it could resist against abnormal packet length, which resulted into the ability to deliver a long message within the distribution of the length on the selected packets. Therefore, with the proposed method, this research, is the first, to associate the length of the message with contain in the payloads and is the first, able to send a long message without adding additional lengths and within the normal packet length's distribution.

2. Previous Work

Packet length is classified as an indirect hidden method because there is no direct modification done to the packets except the data is hidden based on the length of the packets. Padlipsky introduced a packet length covert channel in [19]. Padlipsky associated the length of link layer frames with a symbol to conceal the secret message. Ten years later, Girling demonstrated Padlipsky idea in [20]. Girling represented the length of the link layer with 256 symbols. Which, it required 256 different packet lengths, and each packet length represents bytes of information. The experiment was done in the isolated network to eliminate the noise of other packets such as buffering, reblocking and spurious message insertion from high level protocols. In real networks, the controls of block size and packet length are actually being modulated and depend on the network conditions [23]. Conveniently, Padlipsky method could be very effective within the same network segment [24].

Two decades later, Yao and Zhang in [21] used a secret matrix with 256 rows and randomly associated it with the length of packets. The arrangement of the length in the matrix will be transformed according to the agreement between the sender and receiver. The Yao and Zhang method has successfully improved the Girling. However, a study by Liping in [15] shown that, the randomly packet length will trigger the detection because it produces abnormal network traffic.

To overcome the abnormal network traffic, Liping in [15] proposed a method based on a reference of length. Liping's method required the sender and receiver to agree upon the length of the packets that the sender sent to the receiver. The agreeable length of several packets is set as a default reference. To send a secret message, the sender takes the byte of the message and adds it to the length of the reference. To get the byte of the message, the receiver will deduce the received length with the initial reference. However, as mention by Liping in [22], this method was

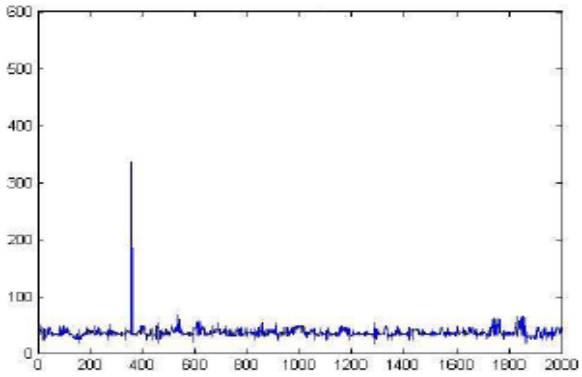


Figure 1: Normal packet length distribution [15]

not efficient when the size of the message is long because the method will update the default reference with every length it received. Therefore, when the message is too long, it will produce abnormal packet length's distribution.

Other noble works on storage covert channels are explained based on the protocol where the covert channels have been exploited.

In [13], Taeshik explained how the IP Identification (IP ID) field being exploited to embed ASCII alphabetic. The method multiplies the ASCII in hex value with 255, since 255x255 is 65535 which just fine to fix into 16 bits fields. However, the method use could trigger suspicious when the same letter in a word occurs. Then Ahsan in [25] improved the method using Toral Automorphism System that used pseudo random sequence to make sure the modified IP identification is random. Yogi Metta's in [6] proposed to exploit the IP ID fields by XOR the byte of the secret message with the IP's version and IP's header length, then, the result will be concatenated with a random number to cover the remaining 8 bits. However, as mention by Murdoch, covert channels with randomly the number in IP ID fields can be detected because it's not by default random [26].

Yogi Metta's in [6] theoretically explain how the value of DF could be use to send a message. The method could successfully implement if and only if we know the MTU of each router. Enrique's in [27] used the IP Offset field to embed the data. The only problems with IP offset field are when the DF is set and there is a data in IP Offset field. This would trigger and IDS or IPS. In [28] Zander demonstrated how TTL is manipulated to send a value 1 or 0. The TTL method is very suitable to send a small amount of data, except. It needs careful study on the variant of Operating System in the network because Fyodor mentions each Operating System use different TTL to identify them. Abad in [29] theoretically described how the Checksum value could carry the data, though; the Windows NIDS

layer will discard the packet if the checksum value is wrong. Moreover, the checksum value will change when the packet goes through the router or NAT.

RFC 792 portrayed Internet Control Message Protocol (ICMP) as a method to help and notified system when an error occurs somewhere in the network path. Most common uses of ICMP are ICMP type 0 (echo reply) and 8 (echo request), which is known as ping. These echo requests and replay could carry 56 bytes of data, and it is ubiquitous among an operating system. On August 1996, Daemon9's demonstrated the ICMP covert channel by exploiting the payload of ICMP type 0 and 8 with malicious data. This is possible because most of the firewall and network perimeter didn't check its payload contents and would allow it to pass [30]. Moreover, the payload of ICMP could carry an arbitrary data [31], therefore, there is numerous ICMP covert channels being exploited and published, such as: Loki [32], ICMP bounce tunnel [33], Ping tunnel [34] and 007Shell [30]. Latest, Zouher in [35] has used ICMP covert channel to send a file and message by exploiting the record router IP header. A lot of ICMP covert channel means the security professional should emphasize their security parameter to limits the ICMP packets. Nevertheless, ICMP will be a great Covert Channel in LAN but not on over the net. This is because most of the firewall will not allow inbound ICMP packet to enter their network [17].Unless used for outbound traffic, ICMP covert channel will be applicable.

Rowland in [11] had shown the basic of TCP covert channel by exploiting the TCP sequence number fields (32 bits). He used the same method as he did for IP Identification, just by multiples the 255x255x255xASCII value. This does not mean Rowland method is naïve, because the method presented is just to shows how the TCP sequence number could be exploited. Ahsan presented method that is more advanced later in [25], where the author encodes the secret message using Toral Automorphism Algorithm. Ahsan's divided the TCP sequence number into two 16 bits. The high 16 bits are used to embed the secret message while the lower 16 bit was generated by the random number. Rutkowska in [36] proposed a more robust method by encrypted the data and XOR it's with one-time-pad key. However, Murdoch specified that, in Rutkowska method. The TCP sequence number didn't exhibit the structure of the TCP ISN as expected in Linux and there is a flaw in the use of DES for encryption, which allows the recovery of the plaintext by statistical information [26]. Therefore, Murdoch later comes out with a more advanced method to show that the covert TCP sequence number look like real TCP sequence generated by Operating System. Murdoch's proposed a method where the data is encrypting with the block cipher that is running in counter mode, which produced different pseudo random sequence for each rekey interval. This 15-bits value than is inserted into the ISNs. The 16-bit field in ISN is set to zero and the rest 15 bits are generated by an RC4 pseudo random number generator [26]. Notes, TCP is the connection oriented therefore the stateful firewall can keep track the TCP state. The exploitation of TCP for covert channel over the net is not viable, unless in LAN because nowadays network perimeter firewall could keep track the TCP connections. Therefore, a single TCP sequence packet will look like a port scan packets.

Despite the concentration on TCP sequence number, Chan in [37], proposed a method call partial acknowledgment to exploit the TCP Acknowledgment number (TCP ACK). Their method is calls partial acknowledgment because the value of TCP ACK is less than ISN +1, as in normal TCP operations. To send a message, let say M, the partial acknowledgment number will be ISN + 1 – M. Therefore, to get back the message, the receiver need to get the value of M by subtract the next ISN+1 – ACK. However, this method is not efficient in the network environment with stateful firewall because the ACK number is less than the ISN + 1. Another problem is, for each secret message, they have to make sure, for each TCP packet, is set with the minimum size of MSS. This will result with a lot of TCP ACK between the sender and receiver despite the OS could handle TCP packets with more payloads.

UDP based on its design principle, is to exchange message with a minimum of protocol mechanism and session management. UDP is connectionless protocol. Therefore, there is no session control to make sure the packets reach the destination. There are few fields could be exploited on UDP for covert channels. The only possible covert channel field on UDP is the source port, and this would be applicable on LAN because the source port will be modified when the datagram goes through the NAT firewall. Conversely, UDP has been used to carry another internet protocol such as IP [38][39] and TCP (Simon) to evade the firewall. Thereby, the covert channel could exits on the protocol on top of UDP stack.

2.1 DNS

DNS or Domain Name System' is used to translate human-readable hostname to numerical IP address and vice versa [40]. In the design, DNS protocol was on top of UDP protocol. This gives advantages to DNS, which, there will be no overhead on the services resource and network perimeter, as there is no connection tracking or session to process. Moreover, there are fields in the DNS protocol, which allows huge bytes to be carried especially in the

query and response [41]. The researcher used DNS as a method to bypass the security perimeter] has exploited these advantages [41. Besides, until now, DNS is less filtered by the organization security perimeter, and this is further proof, when the captive portal allows the DNS to make a query to the internet, although the user hasn't authenticated to the network, which means, the DNS query is independent of the identity of the requestor. This further encouraged the used of DNS query and response, instead, as a simple method to bypass the firewall as a method to tunnel through a network. The used of DNS as a tunnel is the further study by Merlo in [41], which do the comparison on the performance of six DNS tunnels; NSTZ [39], DNSCat, Iodine, TUNS [42], Dns2TCp and OzyManDNS [38].

However, there is a major different between tunnel and covert channels. Generally, tunnel main intentions are just to bypass the security filtering and not to fade the communications [43]. Further, in tunneling, the clients and server have to keep tract of the connections between each other while the connection is active, which, results in high traffic between the communications node [42]. Moreover, in tunnels, the method used to carry their data is not to hide their data appearance as in covert channels. Therefore, the data is being encoded in non-compliant to Base64 encoding [41]. Albeit, the problem highlighted with the tunnels, is not to be highlighted that the tunnel is not good, but to support that, DNS is the good choice for covert channels because, as stated in [41], until now, DNS is the less filtered protocol, which means. It can be used for the purposed covert channels, which meet the indispensability property.

3. The DNS Reference Model

3.1 The step to send and received the message

The study subdivided the entire process into the method similar to OSI model for better understanding as follows:

• Level 0; starting from Alice's side, Clear Message (M) is the readable message that Alice wishes to send to Bob.

• Level 1; Alice's M is encrypted (Em) with a block cipher algorithm and stored the Em in the queue.

• Level 2; The indirect algorithm will associate the corresponding Cm with standard URL name.

• Level 3; The Sp will be injected into the network that will passes through the protection network as normal DNS query packets.

• Level 3; On the Bob side, all received datagram will be picking up and stored in memory stack.

• Level 2; Indirect Detection module will determine the correct Sp. The correct Sp will be processed and the byte of the message will be stored on the stack until the end of the flow control is found.

• Level 1; together with the Sk, the Decryption Decoder will decode Em recover the sent message.

• Level 0 if the Em is successfully decoded, Bob will be able to read the M which sent by Alice.

To be note; for the rationale of the SCCF as in Figure 1, the study assumed the follows conditions:

• Reasoning for the purpose of security. Cm is encrypted using Symmetric encryption algorithm. The Sk is only known to Alice and Bob.

• The objective of this DNS Covert Channel is to hide the secret messages in Sp through Indirect Algorithm.

• The process of the transmission Sp is in sequence order with ideal timing and overt network.

• In some situation, when Sp needs to travel across multiple networks; Sp must not be detected by other nodes. Therefore, Sp must not show any different between normal DNS packets against Sp DNS packets.

• assuming there is no Sp loss because of buffer unavailability or network congested.

3.2 The indirect reference algorithm

The stego method used in DSCCF is based on URL name to represent the Base 16 values. The URL hostname could be any agreeable hostname that is normally used in the network where the sender resides. The preferable solution is to choose the URL that normally requests by the clients in the network. Importantly, the DNS query's datagram should not exceed 300 bytes and 512 bytes for the response datagram as stated in [45]. Albeit, there are certain conditions, which allow the DNS query to specify the response datagram can exceed 512 bytes by using the OPT Resource Record [46]. However, as stated in [42], the length of the URL should not exceed 140 bytes.

The indirect reference module will process the cipher in block size of 4 bits. For each block, the module will find the corresponding Base 16 values. The row of the corresponding Base 16 is the current amount of block being processed. Notes that, the amount of blocks will be mod with 16. As a result, the value will be in between 0 to 15. This value is the row that process will find the position of Based 16 value. Once the value of Base 16 is found in row[n], the module will generate the DNS packets with corresponding URL name and store in stack. The process of covert the Em to corresponding URL name will end when all the byte in the Em has been processed. line.

Table 1

DNS Samples		A	B	C	Average Packet length (%)
Number of Packets		34607	398067	649962	
Time (Minutes)		280	64	121	
Length	40-79	64.82	51.42	50.55	55.6
	80-159	35.18	48.58	49.45	44.4
	160-319	0	0	0	0

4. The Experiment

The proposed indirect algorithm was tested based on the URL name collected from campus network. The proposed schema was compared with [20], [21] and [15] models.

4.1 DNS Dataset

The DNS dataset was collected on-campus network. The study used tcpdump to collect the packets. Table 1 shows the three different samples that have been captured on three different times.

Sample A has about 34,607 thousand standard DNS query packets with average packet's length in between 40 to 79 is about 64.82 percent and 80 to 159 with 35.18 percent. Sample B has about 398,067 thousand standard DNS query packets with average packet's length in between 40 to 79 is about 51.42 percent and 80 to 159 with 48.45 percent. Sample C has about 649,962 thousand standard DNS query packets with average packet's length in between 40 to 79 is about 50.55 percent and 80 to 159 with 49.45 percent.

The most important point with this data, the covert channels will have about 119 packets different lengths to operate and not to exceed 159 packet lengths. Otherwise, its will show an abnormality in the networks. Moreover, it could be tolerated to say that, it should be normal for the covert channels with plus or minus 5 percent of the average packet length in between 40 to 79 and 80 to 159.

4.2 Experiment Setup

The implementation of the proposed covert channel was done using the winpcap library to send and capture the standard DNS query. To capture the packets for the purpose to be validated, evaluated and monitor, the study used Ethereal and Wireshark. The message length of the cipher is 88 bytes. The message was encrypted using 256 block cipher algorithm with shared key. The Oracle Virtual Box was used to running the Windows XP operating systems with 100Mbps.

Table 2

Author's Model		Girling	LAWB	Liping	Propose
Number of Packets		88	88	194	178
Length	40-79	0	12	12.89	60.67
	80-159	58	60	80.41	39.33
	160-319	30	16	13	0

4.3 Experiment Result

In this section, the study discusses the findings based on two comparisons. First, the study compares the statistical of the DNS length to shows the percentages based on the group length as stated in the Table 1. Then, the study compares the distribution of the length as discuss in subsections 4.3.2. Lastly, the study presented the bandwidth that could be achieved.

4.3.1 Packet length comparison

Table 2 shows the results of Girling's, LAWB's, Liping's and propose the schema after successful sending the secret message to Bob.

4.3.1.2 Packet efficiency

Based on the number of DNS packets used to transfer the secret message, the results in Table 2 shown that, the Girling's and LAWB's method, only required 88 packets, which is better than propose and Liping's method. This is because, in Girling's and LAWB's method. The length of DNS packets is directly associated with ASCII's table value. The propose method is better than Liping's method because Liping's method required Alice's to send preliminary DNS packets to Bob as reference length.

4.3.1.2 Data transfer efficiency

Regarding the data-transfer efficiency, Girling's and LAWB was better than propose and Liping's because Girling's and LAWB method can carry one bytes per packets. The propose and Liping's method carries 4 bits of data per packet.

While Liping's and the propose method required more packets than Girling's and LAWB's method, this doesn't mean that Liping's and the propose method is not efficient. Without tempering the packets with secret data, propose and Liping's method was better than timing's covert channel with four bits of data per packet. In timing covert channel, each packet can only transfer one bit of

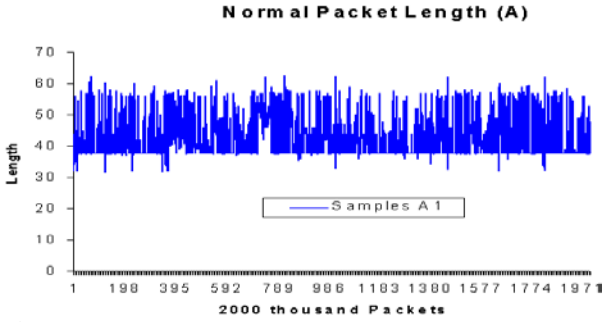


Figure 2

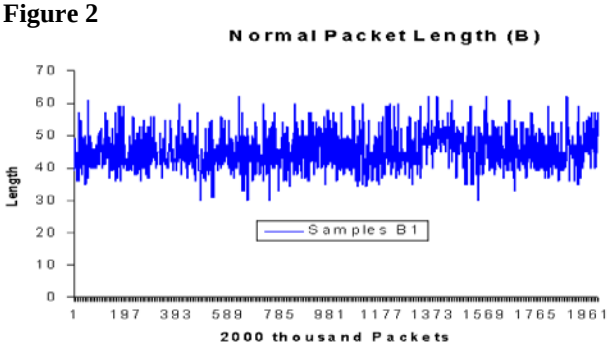


Figure 2

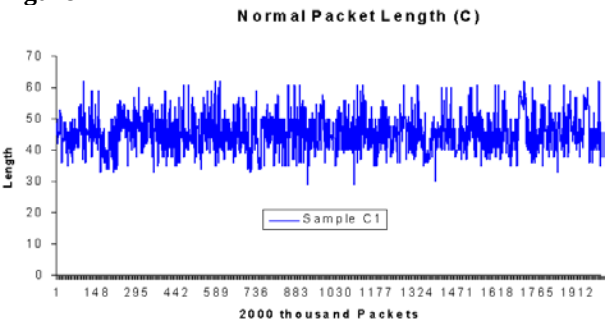


Figure 3

data, which, will require 512 packets to transfer 88 bytes of data.

4.3.1.3 Packet length percentage

The only method where the packet lengths are within the range of 40 to 159 lengths is the propose method. This is because, in the propose method, packet lengths are just a measurement to make sure the covert packets are within the normal packet lengths range. Unlike Girling's, LAWBs and Liping's, where the data was hidden based on the different range of packet lengths, which are bound to the availability of the range of packets in that particular protocol. Liping's method, in [15], based on their test bed HTTP data from Clarknet, was available with only 400 ranges of different HTTP lengths. The limited ranges of length are not limited to Liping's method. Girling's and LAWB was bounded to the same problem where their method could only be used with the range of 55 packet length. Albeit, in HTTP of Clarknet samples, there are 400 packet lengths range. Therefore, the propose method is better, because lengths, is not the limits.

4.3.2 Normal packet length distribution comparison

In normal distribution packet lengths, the study looks into the length of the UDP packets that used to envelop the DNS protocol. The analysis of normal length was done on three DNS dataset as stated in sections 4. Based on the samples, the study analyzes 2000 thousand and 200 packets of each sample and plots a normal distribution of the packet length's graph as in figure 4, 5,6,7,8, and 9. The normal distribution of packet lengths is based on 2000 packets is enough to show the distribution of packets based on a comparison done in [15], however, the 200 normal distribution of packet lengths is required to give the explanation for the results of the experiments based on 88bytes of a cipher message which required 196 packets of DNS to conceal the cipher message.

The normal distribution of packet lengths based on 200 hundred packets as shows in figure 7, 8, and 9 are accordingly with their respective 2000 packets length. Therefore, the two distributions based on 2000 thousand and 200 packets will develop to justify packet length's comparisons. The graph in Figure10 is the packet length's distribution that was plotted based on the result from the propose schema. The propose packet length distributions depict in Figure 10 was compared to the 2000 thousand and 200 hundred normal distributions. The study found that the propose result was normal than Girling's, LAWB's, and Liping's. The Girling, LAWB AND Liping schema is depicted in Figure 11,12, and 13. The result proof that the propose distribution's packet lengths are normal to the normal distributions. The normal distribution is shows in Figure 5 and 8. Therefore, propose indirect method based on Base 16 matrix has been successful.

They should be numbered consecutively throughout the text. Equation numbers should be enclosed in parentheses and flushed right. Equations should be referred to as Eq. (X) in the text where X is the equation number. In multiple-line equations, the number should be given on the last line.

4.3.3 Statistical Test

The study further analyzes to propose packet lengths and the three samples as shown in Table 1 with T-test. The T-test [47], is used to measure significant different in their distributions. The study analyzed the UDP packet lengths on each sample and plot the boxplots to get boundaries of the packet lengths.

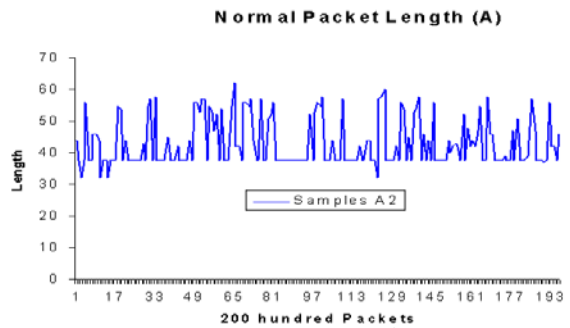


Fig 5: Normal Packet Lengths Distribution of Sample A2

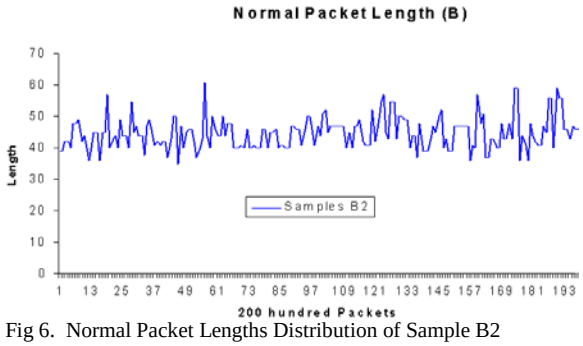


Fig 6. Normal Packet Lengths Distribution of Sample B2

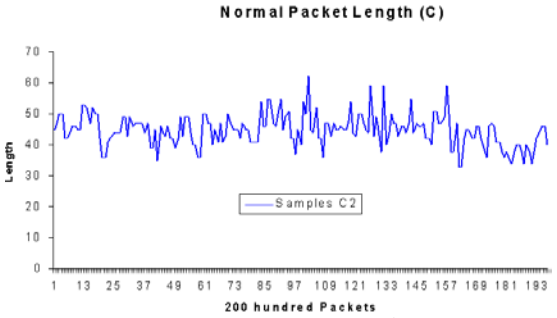


Fig 7. Normal Packet Lengths Distribution of Sample C2

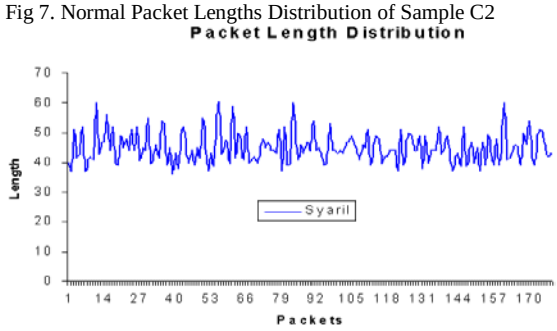


Fig 8. Propose Packet Length Distributions

Based on the boxplots in Figure 12, it could be concluded that the range of the packet length's D (Propose method) didn't same with the range of Packet lengths (A). Subsequently, the hypotheses for the relevant 2-tailed would be of the form:

H: Distribution of D packet lengths didn't same with A.
H: Distribution of D packet lengths is same with A.

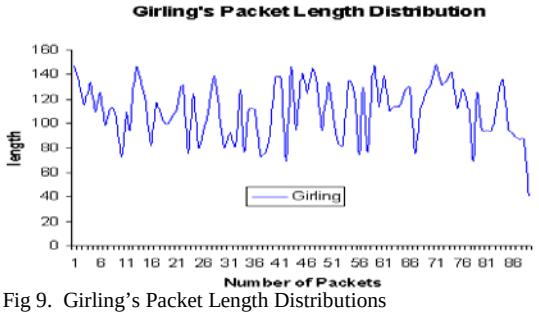


Fig 9. Girling's Packet Length Distributions

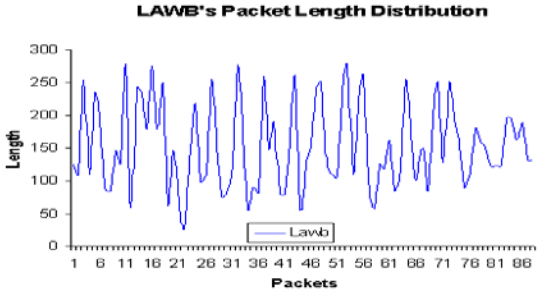


Fig 10. LAWB's Packet Length Distributions

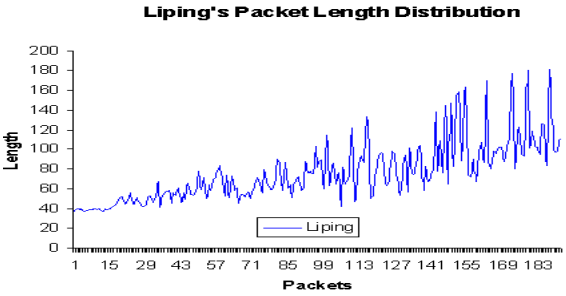


Fig 11. Liping's Packet Length Distribution

The study used Gnumeric statistical software to calculate the tstat, tcrit and p-value, leading to the following conclusions:

As we can expect from Figure 14, there is not enough evidence to reject that the distribution of B and C packet lengths is different from D.. Moreover, based on the p-value for the comparison between C and D, there is sturdy evidence to proof that the distribution of packet lengths between C and D are alike. No less, there is also a correlation between packet lengths for B and D.. Therefore, the t-test has proof that the propose packet lengths covert channel is normal with the campus's packet length distributions.

5 Conclusions

The novel indirect packet length covert channels has been proposed to generated normal packet length which

associated the payload of the packets to more than one symbols as done by previous packet length covert channels. This is done by using a reference matrix of Based 16. The experiment results shows that the covert channel has able to sustain in the upper bound of the average normal DNS packet lengths taken from the campus network and was normal in the packet length distributions which resist the covert channels against abnormal packet lengths observation..

Acknowledgments

This work is supported by National Science Fellowship of Malaysia. Special thanks to Prof Md Dr Mohd Asri Bin Ngadi, Ismail Hamedy and Mohd Nasri Mat Isa.

References

[1] L. Frikha and Z. Trabelsi, "A new covert channel in WIFI networks," Risks and Security of Internet and Systems (CRISIS 08), 2008, p. 255–260.
[2] S. Hamdy, W. El-Hajj, and Z. Trabelsi, "Implementation of an ICMP-based covert channel for file and message transfer," 2008, p. 894–897.
[3] G. Armitage, P. Branch, and S. Zander, "Covert channels in multiplayer first person shooter online games," 2008, p. 215–222.
[4] A. El-Atawy and E. Al-Shaer, "Building covert channels over the packet reordering phenomenon," INFOCOM, 2009, pp. 2186-2194.
[5] A. Desoky, "Listega: List-based steganography methodology," International Journal of Information Security, vol. 8, 2009, pp. 247-261.
[6] Y. Mehta, "Communication over the Internet using Covert Channels," 2005.
[7] J. Lubacz, W. Mazurczyk, and K. Szczypiorski, "Vice over IP," IEEE Spectrum, vol. 47, 2010, pp. 42-47.
[8] G. Fisk, M. Fisk, C. Papadopoulos, and J. Neil, "Eliminating Steganography in Internet Traffic with Active Wardens," Information Hiding, 2002, p. 22.
[9] G. Danezis, "Covert Communications Despite Traffic Data Retention," Security Protocols XVI, B. Christianson, J.A. Malcolm, V. Matyas, and M. Roe, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 198-214.
[10] J. Giffin, R. Greenstadt, P. Litwack, and R. Tibbetts, "Covert Messaging through TCP Timestamps," vol. 2482, 2003, pp. 194-208.
[11] C. Rowland, "Covert channels in the TCP/IP protocol suite. Tech. rep., First Monday," ACM Transactions on Information and Systems Security, vol. 12, 1997, p. Article 22.
[12] S.J. Murdoch and S. Lewis, "Embedding Covert Channels into TCP/IP," The 7th Information Hiding Workshop, 2005, pp. 247-261.
[13] T. Sohn, J. Seo, and J. Moon, "A Study on the Covert Channel Detection of TCP/IP Header Using Support Vector Machine," In Proceedings of the International Conference on Information and Communications Security, 2003, pp. 313-324.
[14] N. Chen, W. Hu, and Z. Xue, "Research of covert channels based on web counters," Shanghai Jiaotong Daxue Xuebao/Journal of Shanghai Jiaotong University, vol. 42, 2008, pp. 1678-1681.
[15] L. Ji, W. Jiang, B. Dai, and X. Niu, "A Novel Covert Channel Based on Length of Messages," International Symposium on Information Engineering and Electronic Commerce, IEEC 2009, 2009, p. 551–554.
[16] H. Khan, M. Javed, S.A. Khayam, and F. Mirza, "Designing a cluster-based covert channel to evade disk investigation and forensics," Computers and Security, vol. 30, 2011, pp. 35-49.
[17] S. Zander, G. Armitage, and P. Branch, "A Survey of Covert Channels and Countermeasures in Computer Network Protocols," Communications Surveys & Tutorials, IEEE, vol. 9, 2007, pp. 44-57.
[18] P.A. Gilbert and P. Bhattacharya, "An approach towards anomaly based detection and profiling covert TCP/IP channels," Proceedings of the 7th international conference on Information, communications and signal processing, Piscataway, NJ, USA: IEEE Press, 2009, pp. 695-699.
[19] M.A. Padlipsky, D.W. Snow, and P.A. Karger, "Limitations of End-to-End Encryption in Secure Computer Networks," Tech. Rep., vol. ESD-TR-78-, 1978.
[20] C. Girling, "Covert Channels in LAN's," Software Engineering, IEEE Transactions on, vol. SE-13, 1987, p. 292–296.
[21] Q. Yao and P. Zhang, "Coverting channel based on packet length," Computer Engineering, vol. 34, 2008.
[22] L. Ji, H. Liang, Y. Song, and X. Niu, "A normal-traffic network covert channel," CIS 2009 - 2009 International Conference on Computational Intelligence and Security, Harbin Institute of Technology, Shenzhen Graduate School, Shenzhen, China: 2009, pp. 499-503.
[23] G. Armitage, P. Branch, and S. Zander, "Error probability analysis of IP Time To Live covert channels," 2007, p. 562–567.
[24] A. Epliremides and S. Li, "A network layer covert channel in ad-hoc wireless networks," 2004, p. 88–96.
[25] K. Ahsan and D. Kundur, "Practical Data Hiding in TCP/IP," Proceedings of the Workshop on Multimedia Security at ACM Multimedia, 2002, pp. 63-70.
[26] S.J. Murdoch and S. Lewis, "Embedding covert channels into TCP/IP," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 3727 LNCS, 2006, pp. 247-261.
[27] E. Cauich, R. Watanabe, C. Science, and A. Zaragoza, "Data Hiding in Identification and Offset IP fields," In Proceedings of 5th International School and Symposium of Advanced Distributed Systems (ISSADS), 2005, pp. 118-125.
[28] S. Zander, G. Armitage, and P. Branch, "Covert Channels in the IP Time To Live Field," In Proceedings of Australian Telecommunication Networks and Applications Conference (ATNAC), 2006.
[29] C. Abad, "IP Checksum Covert Channels and Selected Hash Collision," Technical report, 2001, pp. 1-3.

[30] T. Sohn, J. Moon, S. Lee, D.H. Lee, and J. Lim, "Covert Channel Detection in the ICMP Payload Using Support Vector Machine," Computer and Information Sciences - ISCIS, vol. 2869, 2003, pp. 828-835.
[31] Daemon9, "Project Loki," Phrack, vol. 49, 1996, p. 6.
[32] Daemon9, "Loki2 (the implementation)," Phrack, vol. 51, 1997, p. 6.
[33] Zelenchuk, "Skeeve - ICMP Bounce Tunnel," 2004, http://www.gray-world.net/poc_skeeve.shtml, 2004.
[34] D. Stødle, "ptunnel - Ping Tunnel," 2005, http:// www.cs.uit.no/daniels/PingTunnel, 2005.
[35] S. Hamdy, Z. Trabelsi, and W. El-Hajj, "Implementation of an ICMP-based covert channel for file and message transfer," Proceedings of the 18th International Symposium on Computer and Information Sciences, 2008, pp. 894-897.
[36] J. Rutkowska, "The Implementation of Passive Covert Channels in the Linux Kernel," Proc. Chaos Communication Congress, Dec, 2004.
[37] E. Chan, R. Chang, and X. Luo, "CLACK: A Network Covert Channel Based on Partial Acknowledgment Encoding," IEEE International Conference on Communications, Dresdan: IEEE, 2009, p. 1–5.
[38] D. Kaminsky, "IP-over-DNS using Ozyman," 2004, http://www. doxpara.com/, 2004.
[39] T.M. Gil, "IP-over-DNS using NSTX," 2005, http://thomer. com/howtos/nstx/, 2005.
[40] P. Mockapetris, "DOMAIN NAMES - IMPLEMENTATION AND SPECIFICATION," RFC 1035, IETF, Nov, 1987.
[41] A. Merlo, G. Papaleo, S. Veneziano, and M. Aiello, "A Comparative Performance Evaluation of DNS Tunneling Tools," Computational Intelligence in Security for Information Systems, Á. Herrero and E. Corchado, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 84-91.
[42] P. Neyron, O. Richard, and L. Nussbaum, "On Robust Covert Channels Inside DNS," 24th IFIP International Security Conference, Pafos, Cyprus: 2009, pp. 51-62.
[43] M. Dusi, M. Crotti, F. Gringoli, and L. Salgarelli, "Tunnel Hunter: Detecting application-layer tunnels with statistical fingerprinting," Computer Networks, vol. 53, 2009, pp. 81-97.
[44] F. Petitcolas, M. Kuhn, and R. Anderson, "Information hiding-a survey," Proceedings of the IEEE, vol. 87, 1999, pp. 1062-1078.
[45] D. Hoeflin, K. Meier-Hellstern, and A. Karasaridis, "NIS04-2: Detection of DNS Anomalies using Flow Data Analysis," Global Telecommunications Conference, 2006, pp. 1-6.
[46] P. Vixie, Extension Mechanisms for DNS (EDNS0), 1999.
[47] A. Bhasah, Data Analysis Method, Utusan Publications & Distibutors Sdn Bhd, 2007..

Syaril Nizam Omar is currently a PhD student in the Department of Computer Systems and Communications of the Faculty of Computer Science and Information Systems at the Universiti Teknologi Malaysia. He obtained M.Sc. Information Security from Universiti Teknologi Malaysia (Malaysia) in 2008. He has been involved in lots of academic research since then; presently he is a member of Pervasive Computing Research Group at UTM, while his research interest Covert Channel. He has published in many national and international learned journals.

Ismail Ahmedy is currently persuing PhD in wireless sensor networks at department of computer system and communication, faculty of computer science and information system, universiti teknologi malaysia. He currently hold an academic post in university malaya since 2009. He obtained M.Sc. (Computer Science) from university of queensland, australia in 2009 and B.S.c (Computer Science) from universiti teknologi malaysia in 2006. His research interest is in wireless sensor networks (protocol and signal coverage).

Md Asri Ngadi received his BSc in Computer Science,and the MSc in Computer Systems from Universiti Teknologi Malaysia in 1997 and 1999 respectively, and the PhD degree from Aston University, UK in 2004. He is an associate professor in the Faculty of Computer Science and Information System, Universiti Teknologi Malaysia His research interests are computer systems and security,information assurance and network security.

DNS ID Covert Channel based on Lower Bound Steganography for Normal DNS ID Distribution

Abdulrahman H. Altalhi[1], Md Asri Ngadi[2], Syaril Nizam Omar[2] and Zailani Mohamed Sidek[3]

**[1] Department of Information Technology, College of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia**

**[2] Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia
Skudai, Johor, 81310, Malaysia**

**[3] Information Security Department, Advanced Informatics School (AIS)
Universiti Teknologi Malaysia, Kuala Lumpuer, 54100, Malaysia**

Abstract

The covert channel is a method used to send secret data within a communication channel in unauthorized ways. This is performed by exploiting the weaknesses in packet or network communications with the intention to hide the existence of a covert communication. The DNS identification (DNS ID) method has been exploited by Thyer. However, the major problem in Thyer's implementation is that the encrypted cipher was directly inserted as a DNS ID value, which is abnormal, compared to the normal DNS ID distribution. We have overcome this problem through the application of Steganography to insert the cipher value into the DNS ID. The data set test for normal DNS ID is taken from MAWI. We tested four different message lengths and plotted the distribution graph. We found that the proposed result is normal compared to normal distribution of the DNS ID. Therefore, this method produces a normal distribution for DNS ID covert channel. .

Keywords: *DNS Identification, Covert Channel, Normal Distribution.*

1. Introduction

The covert channel (CC) is a method designed to prevent custodian or network monitoring devices from detecting the information exchanged between two parties. This means, that there should be no way for a warden to observe what is being exchanged between the communicating parties. However, if there is no study performed on the effect of the method used to conceal the secret into the packets, these packets could raise suspicions that will alert the warden. Once the warden has been alerted, the warden may record whatever is being exchanged between the parties and later perform an analysis to grasp that there is secret information in the exchange. This does not mean that the warden has to reveal the information, but it may mean that the warden will be suspicious of the activity.

For example, two parties may be engaged in an exchange of a secret message and protect the information with a cipher. The cipher itself reveals that something important or secret is being transmitted in the exchange – and that could further invoke activity to decrypt the message. In [1], the authors explain the conditions under which the information is more secure via the use of Steganography. Steganography is a method used to conceal the existence of the message in a tangible medium cover, such as a picture, a movie or some music. However, this does not mean the CC is not secure against Steganography. The main difference between CC and Steganography is the medium cover. Steganography hides the message in the file, while CC hides the data in the packets, which are a volatile medium. A packet will be destroyed after it reaches the destination or when it cannot reach its destination. These volatile characteristics have made CC the preferable way to send a secret message.

CC has been used to send malicious messages [2, 3], steal information [4], control a Trojan [5], and leak sensitive information [6, 7]. Albeit, the CC also has good applications, such as protecting anonymity and tracing [7], protecting anonymity and preserving privacy [8] and protecting government information and e-commerce transactions [9].

As of the current research, the physical layer has been exploited up to the application layer for CC [10]. As mentioned in [10], many firewalls have blocked internet traffic to reduce the CC threat. However, as stated in [10, 11], the DNS is less filtered because of the great need for

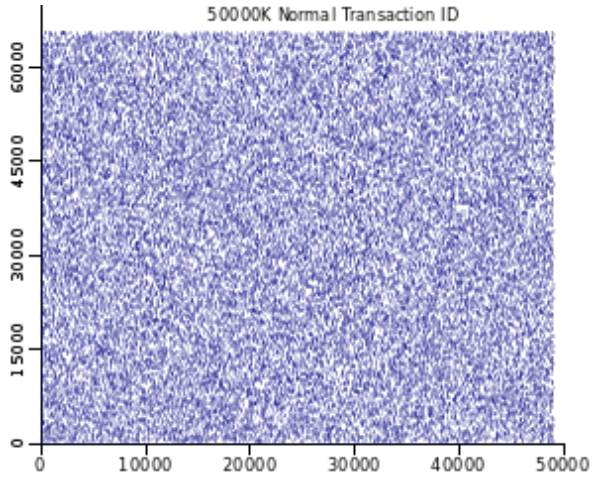


Fig. 1 The normal randomness of the DNS ID distribution taken from MAWI data set.

Internet access. DNS is used to translate the mnemonic name of the server to its corresponding IP Address. DNS is built on top of the UDP protocol, which means that it is connectionless and has low reliability. As reported in [12], based on a study performed on DNS queries for two weeks, the minimum number of queries per second is 11 and the maximum number is 90 queries per second. The high number of queries per second is because duplicated queries are allowed for the purpose of increasing the reliability of the response as required in the relevant RFC. When duplicated DNS queries are received, the DNS server will respond to the queries based on the autonomy field in DNS, which is the DNS ID. The DNS ID can be viewed as an authentication key for each DNS request. As stated in the RFC, the DNS ID should be random enough to make sure that each query has a unique DNS ID.

The unreliability of the DNS protocol and the randomness of the DNS ID give the researcher an opportunity to study the applicability to exploit it for CC. A previous technical report [13] mentions the ability to embed 16 bits of concealed values in the DNS ID. Later, in [11], Thyer elaborates and shows how the DNS ID can be used to send a secret message. Thyer developed a method with a plain insertion and cipher insertion, which prevents the warden from noticing or being able to recognize the hidden information in the DNS ID. However, merely inserting a block cipher string into the DNS ID violates the randomness distribution of the DNS ID. Figure 1 below shows the normal randomness distribution of 50,000 K DNS ID taken from MAWI data set.

For that reason (based on Figure 1), the challenge is not just to craft an encrypted message and embed it into the protocol field but to develop a method that does not violate the property characteristic of the protocol field exploited as shown in [14]. Murdoch shows that the embedded method must adhere to the characteristics of the original design.

Therefore, we would like to present a method based on the characteristic of LSB to embed 8 bits of a secret message into the lower bound or 8 bits of less significant field of DNS ID. We protect the message by encrypting it with a block cipher.

The rest of the paper is organized as follows. Section 2 will discuss the related studies and discussions in the literature review. This is followed by the presentation of the overview of the design in section 3. Section 4 will show how the DNS ID is used to implement the covert channel. And section 5 will discuss the findings and tests. Lastly, the study conclusions are presented and the related improvements are explained in section 6.

2. Related Works and Discussions

The study of the covert channel was originated in 1973 by Lampson. It was then known as the subliminal channel [15]. The first use of a covert channel for a secret purpose was applied when the United States carried out a mission to calculate how many Minuteman missiles they had in a 1000 silos - without revealing which silos actually contained missiles [16]. From 1978 until today, more than a dozen research studies have been performed on covert channels. In this study, we would like to review the work related to covert channels performed on IP, ICMP, TCP, UDP and the Application layer.

2.1 IP Protocol

In [17], Taeshik explained how the IP Identification (IP ID) field can be manipulated to embed ASCII alphabets. This method was used by Rowland to multiply the ASCII as a hex value with 255 because 255x255 is 65535, which is the value of 16 bit fields. The proposed method was excellent in concealing data in the IP ID because the data resembled the value of an IP ID. Note that the initial intention of Rowland was to prove that the IP ID can be exploited to carry a secret message. The design was excellent for sending unique characters. However, the design seemed suspicious if closely monitored, as in the case of the use of duplicate characters. Then, Ahsan in [18] improved the method using a Toral Automorphism System that used a pseudo random sequence to ensure that the modified IP identification is random.

Yogi Metta in [19] theoretically explains how the value of DF could be used to send a message. The method can successfully be implemented if we know the MTU of each router. Cauich and colleagues in [20] used the IP Offset field to embed the data. The only problem with the IP offset field occurs when the DF is set and there is data in IP Offset field. This would trigger an IDS or IPS. In [21], Zander and colleagues demonstrated how the TTL is manipulated to send a value - 1 or 0. The TTL method is

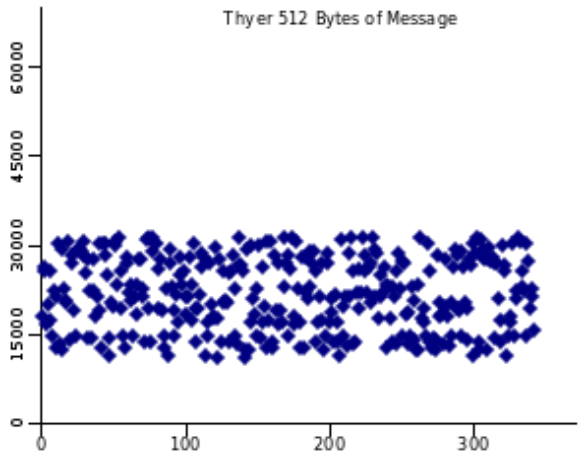


Fig. 2 The randomness distribution of the DNS ID taken from Tyher CC when embedded 512 bytes of a message

very suitable for sending a small amount of data however the variations of Operating Systems in the network need to be carefully studied because Fyodor in [22] mentioned that each OS uses different TTLs to uniquely identify the OS. Abad in [23] theoretically described how the Checksum value could carry the data, although the kernel will discard the packet if the checksum value is wrong. Moreover, the checksum value will change when the packet enters the router. On the other hand, the checksum CC is desirable in a LAN because the detection of checksum CC in a LAN will be very difficult.

2.2 ICMP Protocol

RFC 792 (which describes the Internet Control Message Protocol, abbreviated as ICMP) was designed to help to notify the system in the event that an error occurred in the network path. Its most common use is ICMP type 0 (echo reply) and 8 (echo request). Daemon 9 demonstrated the ICMP covert channel by exploiting the payload of ICMP type 0 and 8. The ICMP payload by default could carry 56 bytes of data. Therefore, the number of ICMP covert channel has increased, such as in Loki [24][25], ICMP bounce tunnel [26], Ping tunnel [27] and 007Shell [28]. Later, Zouher in [29] used an ICMP covert channel to send a file and message by exploiting the record router IP header. The transmission of many ICMP covert channels means the security professional should optimize their security parameters to limit the ICMP packets.

ICMP will be a great CC within the LAN because most of the firewall will not allow inbound ICMP packets to enter their network. Unless it is used for outbound traffic, an ICMP covert channel will be applicable.

2.3 TCP Protocol

Rowland in [30] has shown the basics of a TCP covert channel by exploiting the TCP sequence number (SEQ) fields (32 bits). He used the same method as he did in IP Identification: just multiply the 255x255x255x ASCII value. This multiplication result is fitted within the TCP SEQ field. Again, Rowland's purpose is just to show that the TCP SEQ field is can be exploited for CC. Rutkowska, then, in [31], shows an advanced method by embedding the cipher into the TCP SEQ so that the TCP SEQ field will resemble the normal characteristic of the TCP SEQ field. Later, Murdoch, in [32], shows a better method that resembles the original design of the TCP SEQ field. The Murdoch method fixed the problem in Rutkowska's method by interpreting each TCP SEQ field as an independent field. Therefore, there is no issue when there is no data to be sent.

2.4 UDP Protocol

UDP was designed to exchange messages with minimum protocol overload processing. The only possible covert channel field on UDP is the source port, and it is only applicable on LAN. Conversely, UDP has been used to carry another internet protocol, such as IP [33][34] and TCP. rare.

2.5 DNS

We found that most of the DNS exploits work by bypassing the firewall with the use of a tunnel. DNSTX and DNScat are the tunnels that make use of the DNS query field to carry their data. The DNS query field (as noted in [35]) is used to carry a domain name. The format for the domain name is obvious, so this method is also applicable for sending a secret message over the net. It is not sneaky, as the unusual data in a DNS query is easily detected using a Network Monitor.

Thyer in [11] shows a reasonable level of stealth against the DNS tunnel model. However, the DNS tunnel can be used to send high bandwidth data. Tyher exploits the DNS ID field as the medium carrier to hide the 16 bits of secret data. In essence, the 64 bit cipher is reasonably random. However, the analysis performed with 512 bytes of the message showed that the sub-group of 16 bits from 64-bit ciphers was not randomly distributed. Figure 2 shows the result of the Tyher method in sending 512 bytes of a message. The message was encrypted using the Blowfish encryption algorithm. The figure shows that the secret message was randomly distributed within the range from 11,000 K to 32,000 K. This indeed was different from the normal DNS ID distribution as shown in Figure 1. However, this does not mean that the Thyer CC is easily detected or blocked because the act of embedding the

secret in the DNS ID with a cipher string still produces a good degree of randomness.

Within the DNS header field, we found that the DNS Identification is unique because the field is generated using a pseudo-random method, and the value will not change along the network path until it reaches its destination [36]. Moreover, DNS is carried by UDP. UDP is connectionless, which means there is no tracking mechanism, such as the sequence number or acknowledgement number in TCP.

3. DNS ID CC Design

In this section, we give an explanation of how we design the DNS ID CC that uses the DNS ID as its medium carrier to carry a secret message.

3.1 An Overview of DNS ID CC

To simplify the explanation of the proposed DNS ID CC, we divide the processes into levels or ladders. On the encoder site, the level will start from zero, which indicates the first step or process and will increase subsequently until the embedded process is completed. On the decoder side, the level will start in descending mode until the message is readable to the receiver. In this CC design, we assumed Alice and Bob are communicating openly through the overt network to send their secret messages. The message is protected with the block cipher encryption algorithm. They share a secret key (Sk), which is used to encrypt and decrypt the message. Therefore, with this strategy, the CC design can be explained as follows:

- Level 0: On Alice's side, the message m is the secret message Alice wants to deliver.
- Level 1: Alice encrypts m with the Blowfish algorithm and stores the cipher C in the list.
- Level 2: The CC will divide the C into a sub-group of 8 bits. The sub-group is processed in sequence. The CC will activate the pseudo-random generator and initialize the random seeds. Then, it will generate the random number in the range of 0 to 65535. This random number is the DNS ID that will be used to embed the bytes of the sub-groups. The CC will embed the byte in the lower bound of the DNS ID.
- Level 3: CC will build the DNS packets with the embedded DNS ID. This packet is inserted in the list.
- Level 4: This is the stage where the packet is inserted into the network where the destination is Bob's IP address.
- Level 4: On Bob's side, Bob listens to the DNS port and takes the DNS packets that arrived on a fixed time lapse.
- Level 3: Bob will extract the lower bound DNS ID and store it in the C string. The C string is ready to be decrypted after the complete DNS packet has been received.
- Level 2; The C string will be decrypted with Sk.
- Level 1: The m is ready for Bob. categories

4. Experiment

The experiment that will be used test the proposed CC will result in DNS ID values with the DNS ID from the MAWI data set and the DNS ID values from Tyher model. Four different sizes of a message will be used as the comparisons.

4.1 Dataset Analysis

The DNS data set was based on the MAWI data set captured with tcpdump on a Mac 2008. The size of the tcpdump file after decompressing it is about one gigabyte. We then filter the DNS standard query from the dump file and we obtain approximately 49,056 K of DNS queries. We then further extract the DNS ID of each query and then run the descriptive statistical analysis on the DNS ID. We found that the mean value is 32,778, and the standard deviation (SD) is 18,883. This means that the value is largely dispersed in the range of +/-18883 from the mean. This DNS ID value will be the benchmark to determine whether the proposed method of DNS ID CC is dispersed with the same distribution. Figure 1 shows the distribution of the MAWI DNS ID data set distribution. Notice that the DNS ID is scattered within values from 0 to 65535.

4.2 Experiment Results

Our results will be discussed by showing the graphs of the DNS ID distribution and the Mann-Whitney U test on four different message sizes.

4.2.1 Randomness Distribution Comparison

Our results will be discussed by showing the graphs of the DNS ID distribution and tIn this experiment, we test four message sizes that are in the range of 64 bytes, 128 bytes, 256 bytes and 512 bytes. As mentioned in Section 3.1, the message will be encrypted using the Blowfish encryption algorithm. After encryption, the sizes of the cipher were in the range of 88 bytes, 345 bytes, 689 bytes and 1369 bytes. Figure 3 shows the randomness distribution of the DNS ID based on the result from the proposed CC and Tyher CC. The labels a, c, e, and g are the randomness distribution graphs showing the results of the Tyher CC DNS ID distribution.

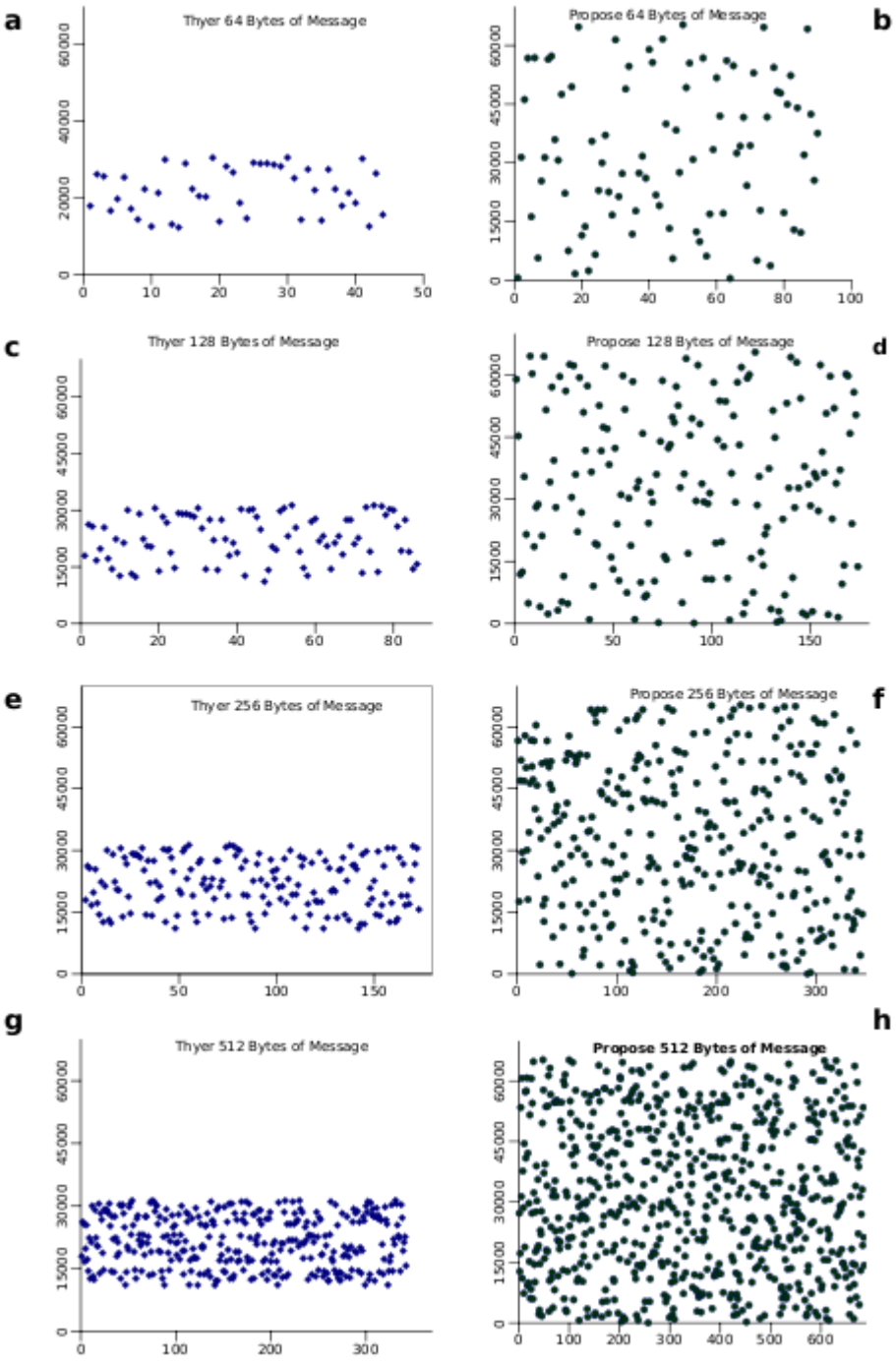


Fig. 3 The result of the DNS ID values based on 4 different messages sizes ranging from 64, 128,256 and 512 bytes. In the left side of the figure, labels a, c, e, and g are the result from the Tyher CC, while labels b, d, f, and h are the result of the DNS ID values from the proposed CC

The labels b, d, f and h show the randomness distributions of our CC DNS ID results. The label a shows the 64 byte message, c the 128 byte message, e the 256 byte message and g the 512 byte message. The label b is for 64 bytes, d for 128 bytes, f for 256 bytes and h for 512 bytes.

Note that the number of the DNS ID is different between the proposed CC and the Tyher CC. In [11], with the Tyher CC design, the embedded method used the entire 16 bit DNS ID field to conceal the message, while we used only the first 8 bits of the DNS ID to conceal the message. For that reason, as shown in the result, for each message size,

the Tyher method only requires half of the packet size compared to what is required by our method.

Based on the results in Figure 3, we can see that for each message size, the randomness distribution of our result is far more dispersed than the Tyher CC. Moreover, our CC result was dispersed in a range that resembled the dispersal range from the MAWI data set. To confirm this, we further investigated the SD values. The SD value for our 64 byte message is 18,727, with a mean value of 32,891. The difference between our mean value and the data set is 133, and the difference with our SD is 156. This proves that, for 64 bytes of a message, though the number of the packets required is doubled, the value of the difference between the proposed method and the data set is trivial compared to the Tyher CC.

We further evaluated the SD and the mean value for the 128, 256 and 512 byte messages of our DNS ID values. For 128 bytes, the SD value is 19,862 and the mean value is 32,432. The SD value for 256 bytes is 18,816 with mean value 32,502. For 512 bytes, the SD value is 18,599 with a mean value of 31,574. We can see that the difference between our mean value and the data set is within the range of +/-1200. Moreover, the difference with the SD is within the range of +/- 1000. Therefore, this proves that the proposed CC DNS ID value is not significantly different from the data set, which is supported by the graphs that show the distribution of the randomness (Figure 3 for the labels b, d, f, and h).

4.2.2 Mann-Whitney U Test

The Mann-Whitney U test is a non-parametric test to test whether the independent sample had an equally large value. In our case, the SD was dispersed widely from the mean value. The results shown in Table 3 further support the analysis we made in Section 4.2.1 and confirmed the results in Figure 3(f) that there is a randomly distribution of the data set test sample. Note that the z-value and the p-value are slightly decreased if compared with the results in Table 2. However, this does not show a significant difference compared with the sample test, as the p-value and the z-value are still high. Albeit, there is a high increase in p-value and z-value for Tyher CC, which show a wider range than the DNS ID of the test data set.

Table 1: The MWU test for 64 bytes of message

64 Bytes	Z-Value	P-Value
Proposed CC	0.48	0.64
Thyer CC	2.59	0.01

Table 2: The MWU test for 128 bytes of message

128 Bytes	Z-Value	P-Value
Proposed CC	0.07	0.94
Thyer CC	2.96	0.003

Table 3: The MWU test for 256 bytes of message

256 Bytes	Z-Value	P-Value
Proposed CC	0.43	0.67
Thyer CC	5.34	0.0001

Table 4: the MWU test for 512 bytes of message

512 Bytes	Z-Value	P-Value
Proposed CC	1.72	0.09
Thyer CC	8.16	0.0001

The result in Table 4 shows a decrease in the p-value and z-value of our DNS ID for 512 bytes of message. However, this is far better than the Thyer CC DNS ID results, which show an increase of more than 30% from the results in Table 3. Overall, the results from the MWU U-tests have shown that the proposed CC was spread within the random distributions of the test data set. Therefore, we can conclude that the MWU U-test results were consistent with the conclusion in sub-section 4.2.1. The MWU test also further supports the DNS ID values we plotted in Figure 3(b, d, f, h), which shows that the DNS ID values are widely spread in the range of 0 to 65535.

5. Conclusions and Future Works

In this paper, we have undertaken studies and implemented a capable DNS ID CC that uses the lower bound of the DNS Transaction ID field to conceal secret values to be transmitted across a network. The solutions we have developed have three main advantages compared to previous studies. First, it can conceal a message inside the DNS ID without violating the random characteristic of the DNS ID. Second, the CC method did not leverage large significant differences from the sample data set, which means it is very difficult for any IDS or IPS to detect that the DNS ID is an object cover that carries a concealed message. Third, the lower bound embedding has successfully proved that it will not affect the normal randomly distribution of the DNS ID.

For the near future, our studies will focus on the need to design a method that can receive and validate CC packets against non-CC packets.

Acknowledgments

The author would like to thank the King Abdulaziz University for supporting the research and National Science Fellowship of Malaysia for the PHD scholarship.

References

[1] Petitcolas, F., Anderson, R., Kuhn, M.: Information hiding-a survey. Proceedings of the IEEE. 87, 1062-1078 (1999).
[2] Trabelsi, Z., Jawhar, I.: Covert File Transfer Protocol Based on the IP Record Route Option. Information Assurance and Security. 5, 64-73 (2010).
[3] Maalej, L., Hammouda, S., Trabelsi, Z.: Towards Optimized TCP/IP Covert Channels Detection, IDS and Firewall Integration. ACSAC. p. 1–5 (2008).
[4] Lewandowski, G., Lucena, N.B., Chapin, S.J.: Analyzing network-aware active wardens in IPv6. ,, Systems Assurance Institute, Syracuse University, Syracuse, NY 13244, United States (2007).
[5] Wang, C., Ju, S.: The new criteria for covert channels auditing. Presented at the (2004).
[6] Moskowitz, I., Newman, R.: Practical Covert Channel Implementation through a Timed Mix-Firewall. Presented at the (2008).
[7] Knight, G.S., Smith, R.: Predictable Design of Network-Based Covert Communication Systems. Presented at the (2008).
[8] Lin, C., Kuo, S., Yarochkin, F., Dai, S., Huang, Y.: Introducing P2P architecture in adaptive covert communication system. First Asian Himalayas International Conference on Internet. pp. 1-7. ,, Kathmandu (2009).
[9] Trabelsi Z., E.H.: Traceroute based IP channel for sending hidden short messages. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 4266 LNCS, 421-436 (2006).
[10] Zander, S., Armitage, G., Branch, P.: A Survey of Covert Channels and Countermeasures in Computer Network Protocols. Communications Surveys & Tutorials, IEEE. 9, 44-57 (2007).
[11] Thyer, J.: Covert Data Storage Channel Using IP Packet Headers, (2008).
[12] Bojan, Z.: Security Monitoring of DNS traffic.
[13] M. Smeets, M.K.: Research Report: Covert Channels, http://www.os3.nl/~mrkoot/courses/RP1/researchreport_2006-02-15_final2.pdf.
[14] Murdoch, S.J., Lewis, S.: Embedding covert channels into TCP/IP. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 3727 LNCS, 247-261 (2006).
[15] Lampson, B.W.: A note on the confinement problem. in Pmc. of the Communications of the ACM, October. 16, 10pp613-615 (1973).
[16] Simmons, G.J.: The History of Subliminal Channels. IEEE Journal on Selected Areas In Communications. 26, 452-462 (1998).
[17] Sohn, T., Seo, J., Moon, J.: A Study on the Covert Channel Detection of TCP/IP Header Using Support Vector Machine. In Proceedings of the International Conference on Information and Communications Security. 313-324 (2003).
[18] Ahsan, K., Kundur, D.: Practical data hiding in TCP/IP. Proceedings of the Workshop on Multimedia Security at ACM Multimedia. 63-70 (2002).
[19] Mehta, Y.: Communication over the Internet using covert channels, https://www.cs.drexel.edu/~vp/CS743/Papers/ypm23-hw2.pdf, (2005).
[20] Cauich, E., Watanabe, R., Science, C., Zaragoza, A.: Data Hiding in Identification and Offset IP fields. In Proceedings of 5th International School and Symposium of Advanced Distributed Systems (ISSADS). 118-125 (2005).
[21] Zander, S., Armitage, G., Branch, P.: Covert Channels in the IP Time To Live Field. In Proceedings of Australian Telecommunication Networks and Applications Conference (ATNAC). (2006).
[22] Fyodor: Remote OS Detection via TCP/IP Fingerprinting. Phrack Magazine. 8, (1998).
[23] Abad, C.: IP Checksum Covert Channels and Selected Hash Collision. Technical report. 1-3 (2001).
[24] Daemon9: Loki2 (the implementation), (1997).
[25] Daemon9: Project Loki, (1996).
[26] Zelenchuk, I.: Skeeve - ICMP Bounce Tunnel, (2004).
[27] Stødle, D.: ptunnel - Ping Tunnel, (2005).
[28] Sohn, T., Moon, J., Lee, S., Lee, D.H., Lim, J.: Covert Channel Detection in the ICMP Payload Using Support Vector Machine. Computer and Information Sciences - ISCIS. 2869, 828-835 (2003).
[29] Hamdy, S., Trabelsi, Z., El-Hajj, W.: Implementation of an ICMP-based covert channel for file and message transfer. Presented at the (2008).
[30] Rowland, C.: Covert channels in the TCP/IP protocol suite. Tech. rep., First Monday. ACM Transactions on Information and Systems Security. 12, Article 22 (1997).
[31] Rutkowska, J.: The implementation of passive covert channels in the Linux kernel. Proc. Chaos Communication Congress, Dec (2004).
[32] Murdoch, S.J., Lewis, S.: Embedding covert channels into TCP/IP. The 7th Information Hiding Workshop. 247-261 (2005).
[33] Kaminsky, D.: IP-over-DNS using Ozyman, (2004).
[34] Gil, T.M.: IP-over-DNS using NSTX, (2005).
[35] Mockapetris, P.: DOMAIN NAMES - IMPLEMENTATION AND SPECIFICATION. RFC 1035, IETF, Nov. (1987).
[36] Bellovin, S.M.: A technique for counting natted hosts. Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment. pp. 267-272. ACM, Marseille, France (2002).

Abdulrahman H. Altalhi is an assistant professor of Information Technology at King Abdulaziz University (KAU). He received a BSc in Computer Science from KAU on December of 1993, MSc on Computer Science from the University of New Orleans on August of 1998. He has obtained his Ph.D. in Engineering and Applied Sciences (Computer Science) from the University of New Orleans on May of 2004. He served as the chairman of the IT department at KAU for two years (2007-2008). Currently, he is the

Vice Dean of the College of Computing and Information Technology of KAU. His research interest include: Networking, Wireless Networks, Computer Security, Software Engineering, and Computing Education.

Md Asri Ngadi received his BSc in Computer Science,and the MSc in Computer Systems from Universiti Teknologi Malaysia in 1997 and 1999 respectively, and the PhD degree from Aston University, UK in 2004. He is an associate professor in the Faculty of Computer Science and Information System, Universiti Teknologi Malaysia His research interests are computer systems and security,information assurance and network security.

Syaril Nizam Omar is currently a PhD student in the Department of Computer Systems and Communications of the Faculty of Computer Science and Information Systems at the Universiti Teknologi Malaysia. He obtained M.Sc. Information Security from Universiti Teknologi Malaysia (Malaysia) in 2008. He has been involved in lots of academic research since then; presently he is a member of Pervasive Computing Research Group at UTM, while his research interest is Information Hiding.

Zailani Mohamed Sidek Received Diploma in Agriculture, University of Agriculture, Malaysia in 1977, BSc in Business Administration from California State University, Fresno, USA in 1982, MSc in MIS from Texas Tech University, USA in 1984, and PhD in Computer Science from Universiti Teknologi Malaysia, Malaysia in 2005. He worked as a Bank Credit Officer in the Agriculture Bank of Malaysia in 1977-1980; Lecturer in the Universiti Teknologi Malaysia, Malaysia in 1982-present; Head of Department in the Faculty of Computer Science & Information Systems, UTM, Malaysia in 1989-1995. He is currently lecturing in the Advanced Informatics School, Universiti Teknologi Malaysia International Campus, Kuala Lumpur, Malaysia.

Integrated Circuit of CMOS DC-DC Buck Converter with Differential Active Inductor

Kaoutar ELBAKKAR[1] and Khadija SLAOUI[2]

[1]Department of physiqcs, University Sidi Mohamed Ben Abdellah, Fès, Morocco

[2]Department of physiqcs, University Sidi Mohamed Ben Abdellah, Fès, Morocco

Abstract

In this paper, we propose a new design of DC-DC buck converter (BC), which the spiral inductor is replaced by a differential gyrator with capacitor load (gyrator-C) implemented in 0.18um CMOS process.

The gyrator-C transforms the capacitor load (which is the parasitic capacitor of MOSFETS) to differential active inductor DAI. The low-Q value of DAI at switching frequency of converter (few hundred kHz) is boosted by adding a negative impedance converter (NIC).

The transistor parameters of DAI and NIC can be properly chosen to achieve the desirable value of equivalent inductance L (few tens µH), and the maximum-Q value at the switching frequency, and thus the efficiency of converter is improved.

Experimental results show that the converter supplied with an input voltage of 1V, provides an output voltage of 0.74V and output ripple voltage of 10mV at 155 kHz and Q-value is maximum (≈4226) at this frequency.

Keywords**:** *DC-DC Buck converter, gyrator-C, differential active inductor, negative impedance converter, quality factor, efficiency.*

1. Introduction

Switching DC-DC converter is ubiquitous in mobile electronic systems. The trend towards low-power dissipation, low voltage, and high accuracy in portable equipments has been driving technology, as well as the parametric requirement of integrated DC-DC converters.

Magnetic theory is at the heart of any non linear regulator with the use of a spiral inductor to transfer energy from input to output in a lossless fashion, and to filter the output from switching signals.

This paper introduces the concept of differential active inductor [1], [2], with high-equivalent inductance value (few µH) and maximum-Q value at switching frequency, thus to allow the complete integration of DC-DC Buck converter, which is especially important in portable power applications.

Section 2 reviews the relevant information of DC-DC buck converter. The realization of the CMOS differential active inductor is presented in section 3, and improved in section 4. The novel concept of buck converter using DAI and NIC is described in section 5. The simulation results obtained on a 1V to 0.74V buck converter are then shown and discussed in section 6. Finally, a conclusion is given in section 7.

2. General view of classical step-down converter in continuous mode

The buck (or step-down) converter is a switching power supply, used to generate a low regulated DC output voltage from higher DC input voltage normally unregulated. [3], [4], [5].

The main components of the BC are a spiral inductor and two switches oppositely phased, that control the storage energy in the inductor, and it's discharging in to the load, Fig.1.

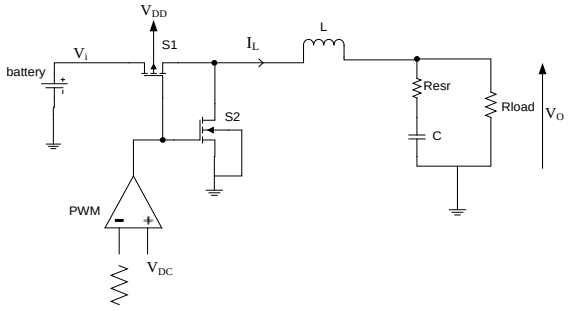


Fig.1 Ideal step down converter.

The switches S_1 and S_2 are usually PMOS and NMOS power transistors driven by a PWM signal at fixed frequency $f_s = \frac{1}{T} = \frac{1}{T_{on}+T_{off}} = \frac{1}{DT+(1-D)T}$, and Varying duty cycle, $D = \frac{T_{on}}{T}$ $0 \le D \le 1$.

The BC operates as follows:

- During T_{on}: switch S_1 is in on-state (closed) and S_2 is in off-state (opened), the inductor is charging, the increase current during T_{on} is giving by:

$$\Delta I_{Lon} = \int_0^{T_{on}} \frac{V_L}{T} dt = \frac{(V_i - V_o)T_{on}}{L} \qquad (1)$$

Where V_i is the input voltage and V_o is the output voltage.

- During T_{off}: switch S_1 is in off-state (opened) and S_2 is in on-state (closed), the voltage a cross the inductor is $V_L = -V_o$, the decrease current during T_{off} is giving by:

$$\Delta I_{Loff} = \int_{T_{on}}^{T_{off}} \frac{V_L}{T} dt = \frac{-V_o}{L} T_{off} \qquad (2)$$

If we assume that the current I_L is the same at $t = nT$ and $t = (n+1)T$, with n an integer.

Therefore:

$$\Delta I_{Lon} + \Delta I_{Loff} = 0$$

$$\Rightarrow \quad \frac{(V_i - V_o)T_{on}}{L} - \frac{V_o}{L} T_{off} = 0$$

$$\Rightarrow \quad V_o = DV_i = \frac{T_{on}}{T} V_i \qquad 0 \le D \le 1$$

The previous study was conducted with the following assumptions:

- The filter capacitor has enough capacitance to keep V_o constant.
- The switches are a very low R_{DSon}.
- Parasitic resistor of inductor is neglected.

3. CMOS differential active inductor

In order to alleviate the limitations imposed on the chip area, and the quality factor (Q) of the spiral inductor, several CMOS active designs were proposed to implement the required on-chip inductance [6], [7].

Fig.2 (a) shows the schematic of the DAI with input $\pm \frac{v_{in}}{2}$ at the source (M_{2a} and M_{2b}) [8], where the pair of stabilizers (M_{3a} and M_{3b}) and negative impedance cross-coupled MOSFET pair (M_{1a} and M_{1b}) have been included at the drain and source of the proposed gyrator circuit respectively. The pair of current sinks M_Q is introducing for external flexible Q tuning, is can be performed by varying I_Q which leads to changes in g_{m1}.

A replica bias circuit M_{L1}, M_{L2} has been introduced to allow current-controlled inductance of the DAI.

Based on a first order small signal analysis, the equivalent RLC circuit of this inductor is shown in Fig.2 (b).

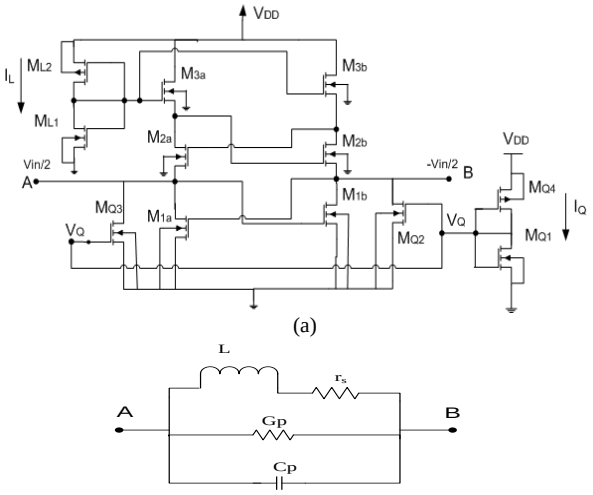


Fig. 2 (a) Differential active inductor, (b) equivalent RLC model of DAI

The input admittance of the DAI is given by:

$$Y_{in} \approx pC_1 + (g_{m2} - g_{m1} + g_{Q3} + g_1 + g_2) + \frac{g_{m2}^2}{(g_{m3} - g_{m2} + g_2 + g_3) + pC_2} \qquad (3)$$

With $C_1 = C_{gs2} + C_{gs1}$

$C_2 = C_{gs2} + C_{gs3} + C_{bd2} + 2C_{gd2}$

Where g_m and g are the transconductance and output conductance of the corresponding transistors.

Neglecting the gate-drain capacitance, we have:

$$L \approx \frac{C_{gs2} + C_{gs3} + C_{bd2}}{g_{m2}^2} \qquad (4)$$

$$r_s \approx \frac{g_{m3} - g_{m2} + g_2 + g_3}{g_{m2}^2} \qquad (5)$$

$$C_p \approx C_{gs2} + C_{gs1} \qquad (6)$$

$$G_p \approx \frac{1}{g_{m2} - g_{m1} + g_1 + g_2 + g_{Q3}} \qquad (7)$$

Based on the RLC model, the resonant frequency of the DAI is given by:

$$f_{res} = \frac{1}{2\pi}\sqrt{\frac{L - r_s^2 C_p - r_s L G_p}{L^2 C_p}}$$

$$\approx \frac{1}{2\pi}\sqrt{\frac{2g_{m2}^2 + g_{m3}(g_{m1} - g_{m2}) - g_{m1}g_{m2}}{C_{gs2}^2 + C_{gs2}(C_{gs1} + C_{gs3})} - \frac{(g_{m3} - g_{m2})^2}{(C_{gs2} + C_{gs3})^2}} \qquad (8)$$

- If the frequency f is much lower than the resonant frequency f_{res}, the RLC model will become inductive.

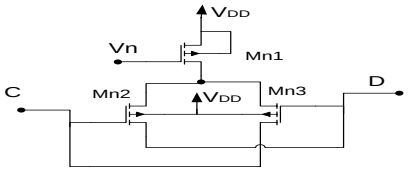
The quality factor of DAI is defined as the ratio of the imaginary part to the real part of input impedance of DAI:

$$Q_0 \Big/ \omega < \omega_{res} = \frac{L\omega}{r_s(G_p r_s + 1) + L^2 G_p^2 \omega^2} \approx \frac{(C_{gs2} + C_{gs3}) g_{m2}^2 \omega}{(g_{m3} - g_{m2})(g_{m3}(g_{m2} - g_{m1}) + g_{m1} g_{m2}) + (C_{gs2} + C_{gs3})^2 (g_{m2} - g_{m1}) \omega^2} \quad (9)$$

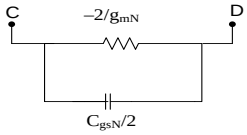
Unfortunately, this structure of DAI doesn't exhibits high-Q at switching frequency (≈100 Khz) [8]

4. Q-enhancement of active inductor

The low Q value at medium frequency can be boosted by adding negative impedance converter (NIC).
Fig.3 shows a simple NIC circuit, it's a cross connected differential pair [9],[10].



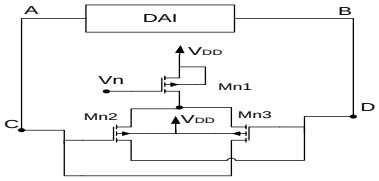
(a)



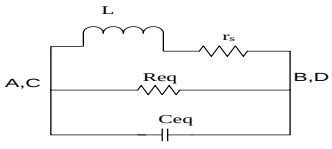
(b)

Fig.3 (a) Schematic of simple NIC. (b) Small signal equivalent circuit.

If we assume that the two transistors M_{n2} and M_{n3} are the same size, the negative differential resistance is $-2/g_{mN}$ (can be tuned by V_n), and the parallel capacitance is $C_{gsN}/2$.
By parallel connecting it to the DAI as shown in Fig.4 (a), and the RLC model equivalent is shown in Fig.4 (b).



(a)



(b)

Fig. 4 (a) Schematic of DAI with high-Q at medium frequency. (b) RLC model equivalent circuit.

With:

$$C_{eq} = C_p + \frac{C_{gsN}}{2}$$

$$R_{eq} = R_p \,//\, \frac{-2}{g_{mN}}$$

The self-resonant frequency becomes:

$$f_{res} = \frac{1}{2\pi}\sqrt{\frac{L - r_s^2 C_{eq} - r_s L G_{eq}}{L^2 C_{eq}}} = \frac{1}{2\pi}\sqrt{\frac{G_m}{(C_{gs2} + C_{gs3})(2C_{gs1} + 2C_{gs2} + C_{gsN})} - \frac{(g_{m3} - g_{m2})^2}{(C_{gs2} + C_{gs3})^2}} \quad (10)$$

With

$$G_m = 2g_{m2}(2g_{m2} - g_{m1}) + g_{mN}(g_{m3} - g_{m2}) + 2g_{m3}(g_{m1} - g_{m2})$$

We operated at frequency much lower than the resonant frequency.
The Q-enhancement value is:

$$Q_{enh}(\omega) \Big/ _{\omega << \omega_{res}} = \frac{R_{eq} L\omega}{r(r + R_{eq}) + L^2\omega^2} = \frac{(C_{gs2} + C_{gs3}) g_{m2}^2 \omega}{(g_{m3} - g_{m2})(g_{m3} g_{m2} - (g_{m1} + \frac{g_{mp2}}{2})(g_{m3} - g_{m2})) + (C_{gs2} + C_{gs3})^2 (g_{m2} - g_{m1} - \frac{g_{mp2}}{2})\omega^2} \quad (11)$$

To maximize Q_{enh} at switching frequency, the transistor parameters can be properly chosen such that the negative resistance of the NIC ($-2/g_{mN}$) compensates for the loss from G_p and r_s at the frequency interest.

$$real\left[\frac{1}{r_s + jL\omega}\right] + G_p - \frac{g_{mN}}{2} = 0$$

As a result, a peak Q factor can be achieved at:

$$\omega_{Q_{max}} = \frac{1}{C_{gs2} + C_{gs3}} \sqrt{\frac{(g_{m3} - g_{m2})(g_{m3}^2 - 2g_{m3} g_{m2})}{g_{m2} - g_{m1} - \frac{g_{mp2}}{2}}} \quad (12)$$

To optimize the efficiency of BC, the transistor parameters can be chosen such that the peak Q factor frequency is few hundred kHz (switching frequency).

5. Model of Buck converter with a high Q differential active inductor:

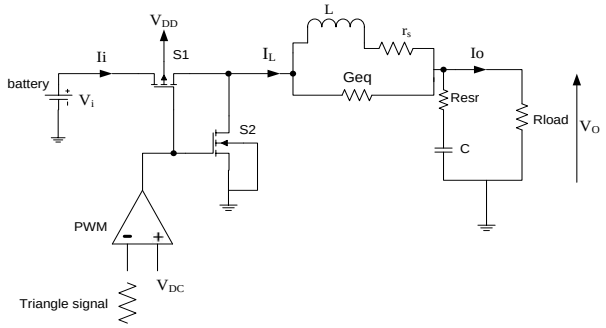


Fig.5 Buck converter with differential active inductor.

Fig.5 shown a BC with DAI and NIC, operated at frequency lower that resonant frequency [11].
The current through the active inductor can be expressed as:

$$I_L = I_{D1} - I_{D2} \qquad (13)$$

In terms of the on duty cycle is:

$$I_L = \frac{I_{D1}}{D} = \frac{I_{D2}}{1-D} \qquad (14)$$

I_L equal to the dc output current I_o, I_{D1} is the dc component of the first switch S_1 current (equal to the dc input current I_i), I_{D2} is the dc component of the second switch S_2 current.
The switches current are:

$$i_{D1} = \begin{cases} \frac{\Delta i_L}{DT} t + I_L - \frac{\Delta i_L}{2} & For\ 0<t \le t_{on} \\ 0 & For\ t_{on} < t \le T \end{cases} \qquad (15)$$

$$i_{D2} = \begin{cases} 0 & For\ 0<t \le t_{on} \\ -\frac{\Delta i_L}{(1-D)T}(t-DT) + I_L + \frac{\Delta i_L}{2} & For\ t_{on} < t \le T \end{cases} \qquad (16)$$

Where: $\Delta i_L = A I_L \frac{(1-D)}{L f_s}$ is the peak-to-peak ripple current of the inductor, and $A = \frac{R_N R_P R_L}{R_L(R_N + R_P) + R_N R_P}$

The rms values of the switches current are obtained as:

$$I_{D1rms} = \sqrt{\frac{1}{T}\int_0^{t_{on}} i_{D1}^2 dt} = I_L\sqrt{D(1+k_I^2)} \qquad (17)$$

$$I_{D2rms} = \sqrt{\frac{1}{T}\int_{t_{on}}^{T} i_{D2}^2 dt} = I_L\sqrt{(1-D)(1+k_I^2)} \qquad (18)$$

Where: $k_I = \frac{\Delta i_L}{\sqrt{12} I_L} = A\frac{(1-D)}{\sqrt{12} f_s L}$ (19)

The power loss in the MOSFETS is found as:

$$P_{r_{ds1}} = r_{ds_1} I_{D1}^2 \frac{(1+k_I^2)}{D} \qquad (20)$$

$$P_{r_{ds2}} = r_{ds_2} I_{D2}^2 \frac{(1+k_I^2)}{(1-D)} \qquad (21)$$

As r_{DS1} and r_{DS2} are the MOSFET on- resistance.
The rms value of the active inductor current is:

$$I_{Lrms} = \sqrt{\frac{1}{T}\int_0^T i_L^2 dt} = I_L\sqrt{(1+k_I^2)} \qquad (22)$$

The power loss in the active inductor is obtained:

$$P_L = \frac{r_s R_{éq}}{r_s + R_{éq}} I_{Lrms}^2 \qquad (23)$$

The current through the filter capacitor is approximately equal to the ac component of the inductor current and is given by:

$$ic = \begin{cases} \frac{\Delta i_L}{DT} t - \frac{\Delta i_L}{2} & for\ 0<t \le DT \\ -\frac{\Delta i_L}{(1-D)T}(t-DT) + \frac{\Delta i_L}{2} & for\ DT < t \le T \end{cases} \qquad (24)$$

The rms value of the capacitor current is:

$$I_{crms} = \sqrt{\frac{1}{T}\int_0^T i_c^2 dt} = k_I I_L \qquad (25)$$

The power loss in the filter capacitor is:

$$P_{Resr} = R_{esr} I_{crms}^2 = R_{esr} k_I^2 I_L^2 \qquad (26)$$

One can estimate the efficiency of the buck converter:

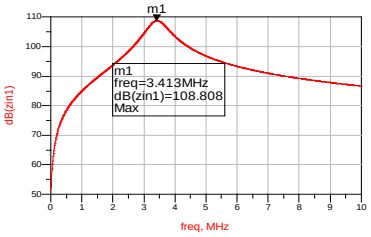
$$\eta = \frac{V_o I_o}{V_i I_i} = \frac{1}{1 + \frac{[Dr_{DS1} + (1-D)r_{DS2} + r_L](1+k_I^2) + R_{esr}k_I^2}{R_{load}}} \qquad (27)$$

As:

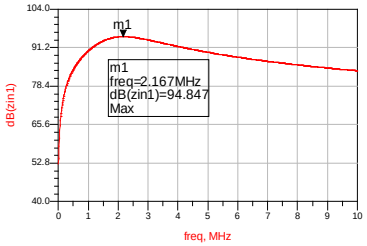
$$r_L = \frac{r_s R_{éq}}{r_s + R_{éq}} .$$

6. SIMULATION RESULTS

The DAI was simulated in 0.18um CMOS technology. Fig.6 shows the variation of input admittance versus frequency of the DAI. Current dissipation of the DAI is 14uA, DAI without NIC resonance at 3.41MHz, and the DAI with NIC resonance at 2.16MHz.



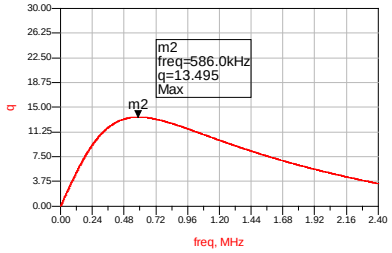
(a)



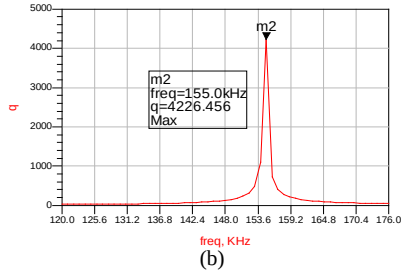
(b)

Fig.6. Variation of input admittance of the DAI: (a) without NIC. (b) With NIC.

Fig.7 shows the quality factor versus frequency of the DAI. Where the DAI with NIC presented a maximum quality factor Qd≈4226 at 155 KHz.



(a)

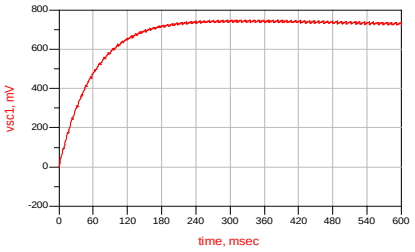


(b)

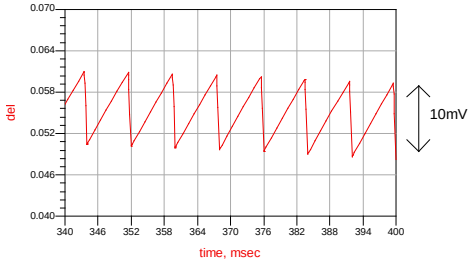
Fig.7 Quality factor of the DAI: a) without NIC. (b) With NIC.

The BC with DAI and NIC is supplied with an input voltage of 1V and switching frequency of 155 kHz.
Fig.8 (a) represents the output voltage of this BC, and the output ripple voltage is shown in Fig.8 (b).
Thus the ripple of output voltage is expressed as follows [5]: $$\Delta V_o = A\frac{(V_i - V_o)D}{16LCf_s^2} = \frac{\Delta i_L}{8Cf_s}$$

and the efficiency of this BC equal: $$\eta = \frac{V_o I_o}{V_i I_i} = 0.75$$



(a)



(b)

Fig.8 Buck converter with differential active inductor: (a) Output voltage of the Buck converter. (b) Output ripples voltage of 10mV.

Fig.9 is the transient response of output voltage with variation of capacitance in the LP filter. The recovery time is in the order of 30ms for capacitance C=100nF and 180ms for C=1uF.

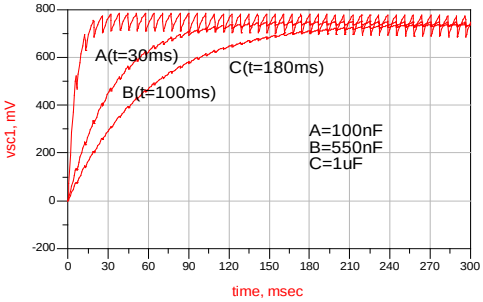


Fig.9 Transient response of output voltage with variation capacitance in the LP filter.

7. Conclusion

In this paper we simulated a buck converter implemented in a 0.18um technology with CMOS differential active inductor paralleled with active negative resistor which uses a minimum number of transistors and presented a high quality factor. The measurement results show that the Buck converter is supplied with an input voltage of 1V and switching frequency of 155 kHz, an low power consumption (14uW), an output voltage of 740mV and output ripple voltage of 10mV.
Due to the use of active inductors, no distributed elements or spiral inductors are required. A significant reduction in chip area can be achieved.
The present work proves that the DAI aren't only suitable for RF applications, but also at medium frequency.

REFERENCES

[1] H. Xiao, R. Schaumann, "Very-High-Frequency Low Pass Filter Based on a CMOS Active Inductor", IEEE International Symposium on Circuit and Systems (ISCAS2002), May 2002.
[2] Liang-Hung Lu, Member, IEEE, Hsieh-Hung Hsieh, Student Member, IEEE, and Yu-Te Liao, "A Wide Tuning-Range CMOS VCO With a Differential Tunable Active Inductor", IEEE TRANSACTIONS ON MICROWAVE THEORY AND TECHNIQUES, VOL. 54, NO. 9, SEPTEMBER 2006, pp 3462- 3468.
[3] John Tucker, "Understanding output voltage limitations of DC/DC buck converters", *Analog Applications Journal* 2008.
[4] Mariko Shirazi, Student Member, IEEE, Regan Zane, Senior Member, IEEE, and Dragan Maksimovic, Senior Member, IEEE, "An Autotuning Digital Controller for DC–DC Power Converters Based on Online Frequency-Response Measurement", IEEE TRANSACTIONS ON POWER ELECTRONICS, VOL. 24, NO. 11, NOVEMBER 2009, pp. 2578-2588.

[5] Volkan Kursun, Siva G. Narendra, Vivek K. De, and Eby G. Friedman, "Analysis of Buck Converters for On-Chip Integration With a Dual Supply Voltage Microprocessor", IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS, VOL. 11, NO. 3, JUNE 2003, pp. 514-522.
[6] Hsieh-Hung Hsieh, Student Member, *IEEE*, Yu-Te Liao, Liang-Hung Lu, Member, IEEE, "*A* Compact Quadrature Hybrid MMIC Using CMOS Active Inductors*"*, IEEE TRANSACTIONS ON MICROWAVE THEORY AND TECHNIQUES, VOL. 55, NO. 6, JUNE 2007, pp 1098-1104.
[7] M. Grozing, A. Pascht, M. Berroth, "A 2.5V CMOS differential active inductor with tunable L and Q for frequencies up to 5GHZ", International microwave symposium digest, VOL. 1, May 2001, pp. 575-578.
[8] Chun-Lee Ler, Student Member, IEEE, Abu Khari bin A'ain, *Member*, *IEEE*, and Albert V. Kordesch, Senior Member, IEEE, "Compact, High-Q, and Low-Current Dissipation CMOS Differential Active Inductor", IEEE MICROWAVE AND WIRELESS COMPONENTS LETTERS, VOL. 18, NO. 10, OCTOBER 2008, pp. 683-685.
[9] Karim Allidina and Shahriar Mirabbasi, "A Widely Tunable Active RF Filter Topology", IEEE International Symposium on Circuits and Systems, 2006, pp. 879-88.
[10] Yue Wu and Mohammed Ismail," A Novel CMOS Fully Differential Inductorless RF Bandpass Filter", IEEE International Symposium on Circuits and Systems, May 28-31, 2000, pp. 149-150.
[11] Dariusz Czarkowski, Marian K. Kazimierczuk, "Energy-Conservation Approach to Modeling PWM DC-DC converters", IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS, VOL. 29, NO. 3, JULY 1993, pp 1059-1063.

Kaoutar ELBAKKAR: PhD candidate at the Faculty of sciences DHAR ELMAHRAZ of Fez. She held an Extended Higher Studies Diploma of the Faculty of Sciences at Fez in 2009 by working on modeling by fuzzy logic.

Khadija SLAOUI: Professor of microelectronics, digital devices, and VHDL at the Faculty of Sciences, Fez, Morocco. She received from the INPT Toulouse France his Doctorate 3rd cycle degree 1984, and state Doctorate degree in 2000.

Improving Security Levels of IEEE 802.16e Authentication By Diffie-Hellman Method

Mohammad Zabihi[1] , Ramin Shaghaghi[2], Mohammad Esmail kalantari[3]

[1] Department of Electrical Engineering, Islamic Azad University, Shahre- Rey Branch, Tehran, Iran

[2] Department of Electrical Engineering, Islamic Azad University, Shahre- Rey Branch, Tehran, Iran

[3] Department of Electrical Engineering, Islamic Azad University, Shahre- Rey Branch, Tehran, Iran

Abstract

In this paper, we proposed an authentication method according to Diffie-Hellman. First, we introduce different methods for authentication in IEEE.802.16 then we proposed an authentication method according to Diffie-Hellman and in the last we compare different methods for authentication to improve security in IEEE802.16e. CPN is a useful for simulation and compare protocol together so we use CPN tools in this paper..

Keywords: *wimax, authentication, color petri net, pkm, diffie hellman.*

1. Introduction

The importance of IEEE 802.16, Worldwide Interoperability for Microwave Access (Wimax) is growing and will complete with technologies such as 3G. The acceptance and adoption of technologies also depend on security. IEEE 802.16 Wimax standard consists of a protocol stack with well-defined interfaces. The Wimax protocol layer contains MAC layer and PHY layer. MAC layer includes three sub-layers contain of: The Service Specific Convergence Sub-layer (MAC CS), the MAC Common Part Sub-layer (MAC CPS) and the Security Sub-layer or Privacy Sub-layer. The former IEEE 802.16 standards used the Privacy and Key Management (PKM) protocol which had many critical drawbacks. In IEEE 802.16e, a new version of this protocol called PKMv2 is released. The authentication and key management protocols are specified in the security sub layer of IEEE 802.16 standard. The security sub layer is meant to provide subscribers with privacy and authentication and operators with strong protection from theft of service. Authentication options are: unilateral authentication, mutual authentication and no authentication sections IEEE S02.16e-2005 standard states that PKM has two versions PKMvl and PKMv2, and it allows for four types of authentication":

- A.RSA base authentication-PKI system(public key infrastructure).
- EAP based authentication (optional).
- RSA based authentication followed by EAP authentication.
- Diffie Hellman base authentication.

2. RSA Base authentication protocol

2.1 Pkmv1 Authentication Protocol

SS uses Message 1, formally named as the Authentication Information Message, to push its X.509 certificate which identifies its manufacturer to BS. BS uses this certificate to decide whether SS is a trusted device. BS may use this message in order to allow access only to devices from recognized manufacturers, according to its security policy. SS sends Message 2, named as the Authorization Request immediately after Message 1(figure.1). Message 2 consists of SS's X.509 certificate with the SS public key, its security capabilities which are actually the authentication and encryption algorithms that SS support, and the security association identity (SAID) which is the id of the secure link between SS and BS. Using the certificate, BS determines whether to authorize SS; and the public key of SS which is also in the certificate lets BS construct Message 3 [1]. If successful, namely SS is authorized after BS verifies its certificate, BS responds with Message 3, the Authorization Reply. This message includes the AK, encrypted using the RSA public-key encryption protocol using the public-key of SS which was obtained in the previous message, the lifetime of the AK as a 32-bit

unsigned number in unit of seconds, the sequence number for AK as a 4-bit value and the list of SA descriptors each including an SAID and the SA cipher suit [1].

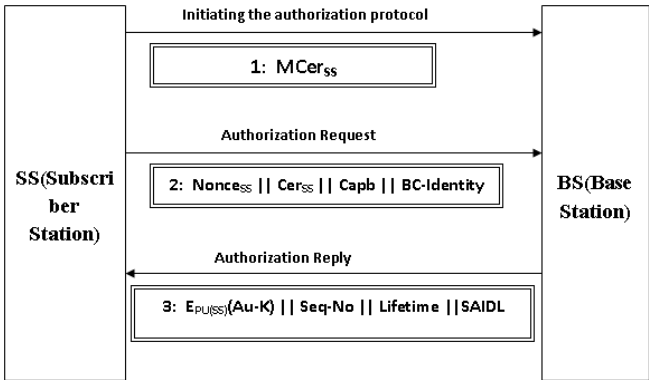


Fig. 1 pkmv1 authentication protocol.

2.2 Pkmv2 Authentication Protocol

The latest standard, IEEE 802.16e-2005, includes a new version (PKMv2) of the protocol that caters for the shortcomings of the first version. PKMv2 supports two different mechanisms for authentication: the SS and the BS may use RSA-based authentication or Extensible Authentication Protocol (EAP) -based authentication. We will focus in this paper on RSA based authentication for PKMv2 authentication protocol. The flow of messages exchange in RSA-based authentication is shown as follows(figure.2): The SS initiates the RSA-based mutual authentication process by sending two messages. The first message contains the manufacturer X.509 certificate. The second, authorization request message, contains the SS's X.509 certificate, 64-bit SS random number Ns, list of security capabilities that the SS supports, the SAID and the SS signature. If the SS is authenticated and authorized to join the network, the BS sends an authorization reply message. In the response message, the BS includes the 64-bit SS random number Ns received, its own 64-bit random number Nb, a 256-bit key pre-primary authorization key (pre-PAK) encrypted with the SS's public key, the pre-PAK key lifetime and its sequence number, a list of SAIDs (one or more), the BS's X.509 certificate and BS's signature in the authorization reply. The SS verifies liveness by comparing the Ns it sent with the received Ns in the authorization response message. It then extracts the PAK, because only the authorized SS can extract the PAK. This can be used as a proof of authorization. Finally, the last message of this authentication is send by the SS to confirm the authentication of the BS. The SS includes the BS random number Nb received in the authorization response message, used to proof liveness, the SS's MAC address and a cryptographic checksum of the message. At the end of the RSA authorization exchange, both SS and BS are authenticated by each other [5].

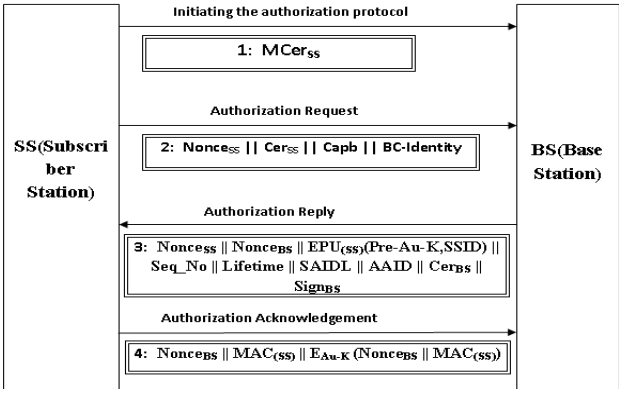


Fig. 2 pkmv2 authentication protocol.

the Extensible Authentication Protocol (EAP) is an authentication framework that is widely used in WiFi/802.11 and Wimax/ 802.16 wireless networks. EAP is a basis to transfer authentication information between a client and a network. It provides a basic request/response protocol framework over which to implement a specific authentication algorithm, so called EAP method. Commonly used EAP methods are EAP-MD5, EAPLEAP, EAP-TLS, EAP-TTLS and EAP-PEAP. Within the EAP framework, three entities are involved in the authentication process: Supplicant, Authenticator, and Authentication Server. The supplicant is a user that is trying to access the network. It is also known as the peer. The authenticator is an access point (AP) that is requiring EAP authentication prior to granting access to a network. It provides users a point of entry into the network. The authentication server (AS) is the entity that negotiates the use of a specific EAP method with an EAP supplicant, then validates the supplicant, and authorizes access to the network.

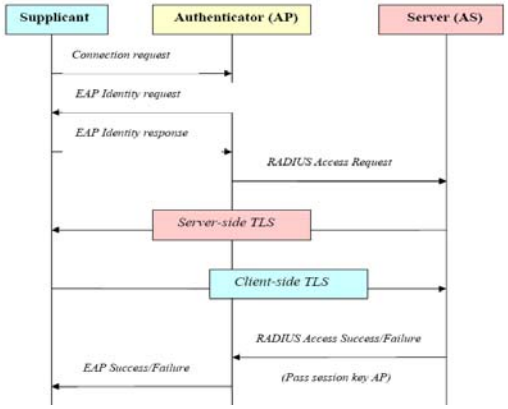


Fig. 3 EAP TLS base authentication.

Typically, the supplicant is a mobile station (MS) and the authentication server is a Remote Authentication Dial-In User Service (RADIUS.Figure.3 shows the brief message flow of EAP-TLS in a WLAN network. The AP creates a RADIUS Access Request using the supplicant's identity and sends it to the AS.The AS then provides its certificate to the supplicant and asks for the supplicant's certificate. The supplicant provides its certificate to AS if the received AS's certificate is valid. After the AS validates the supplicant's certificate, it will send the result message RADIUS Access Success/Failure to deny or permit access to the network [3].

4. Proposed Protocol

Diffie-Hellman key exchange (D-H) is a cryptographic protocol that allows two parties that have no prior knowledge of each other to establish together a shared secret key over an insecure communications channel. Then they use this key to encrypt subsequent communications using a symmetric-key cipher. The scheme was first published publicly by Whitfield Diffie and Martin Hellman in 1976. the Diffie- Hellman exchange by itself does not provide authentication of the communicating parties and is thus susceptible to a man-in-the-middle attack. An attacking person in the middle may establish two different Diffie-Hellman key exchanges, with the two members of the party "A" and "B", appearing as "A" to "B", and vice versa, allowing the attacker to decrypt (and read or store) then re-encrypt the messages passed between them. A method to authenticate the communicating parties to each other is generally needed to prevent this type of attack [2]. As shown in Figure.4, AS sends a request message to the BS that includes the certificate. Then the AS responds this message by sending (Cert BS,P(nonce),H(Ybs||f(nonce)) to A. Being a RSA encryption, P can encrypt nonce and User A can decrypt to receive nonce.

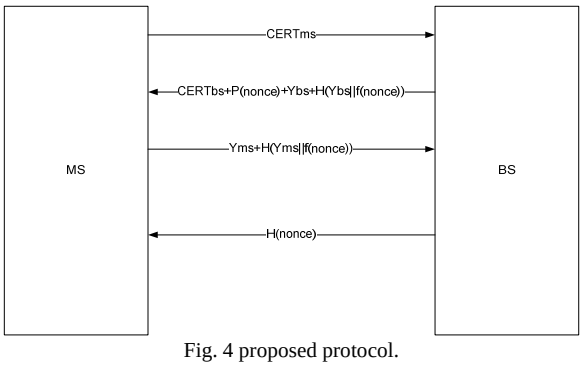


Fig. 4 proposed protocol.

Now, nonce is shared by MS and BS. And then, AS sends (Yms, $H(Y_{ms}$ ||nonce)) to BS, BS calculates H '(Y_{AS} ||nonce). If H '(Y_{AS} ||nonce) is equal to H(Y_{AS} ||nonce), BS believes this message sent by AS, or interrupts this communication. Similarly, AS calculates H '(Y_{BS} ||f(nonce)) by Y_{BS} and from AS. If H '(Y_{BS} ||f(nonce)) is equal to H (Y_{BS} ||f(nonce)), AS believes this message sent by BS and calculates the K= $(Y_A)X_B$ mod q, or interrupts this communication. After A sends a massage H(nonce) as a confirmation signal to BS, AS and BS calculate the:

$$AK= (Y_{AS})X_{BS} \bmod q=(Y_{BS})X_{AS} \bmod q \qquad (1)$$

5. Threat analysis

intercepts messages during the process of communication establishment or a public key exchange and then retransmits them, tampering the information contained in the messages, so that the two original parties still appear to be communicating with each other. In a man-in-the-middle attacks, the intruder uses a program that appears to be the (access point) AP to SS and appears to be the SS to AP. Denial of Service (DOS) attack is an incident in which a subscriber is deprived of the service of a resource they would normally expect to have. A considerable amount of denial of service attacks implement across the Internet by flooding the propagation medium with noise and forge messages. The victim is overwhelmed by the sheer volume of traffic, with either its network bandwidth or its computing power exhausted by the flood of information. Almost all the DOS vulnerabilities in Mobile wimax standard are due to unauthenticated or unencrypted management messages. We discussed these vulnerabilities in three processes: the initial network process, resource saving process and handover process[7]. eavesdropping of management messages is a critical threat for users and a major threat to a system. For example, an attacker could use this vulnerability to verify the presence of a victim at its location before perpetrating a crime. Additionally, it might be used by a competitor to map the network. Another major vulnerability is the encryption mode based on data encryption standard (DES). The 56 bit DES key is easily broken with modern computers by brute force attack. Furthermore, the DES encryption mode includes no message integrity or replay protection functionality and is thus vulnerable to active or replay attacks. The secure AES encryption mode should be preferred over DES. Eavesdropping mostly affects the transfer of information and rarely causes system outage. The assessment of the eavesdropping threat is minor to the system but high for the user [8] .in table.1 we show vulnerability for each method. the proposed authentication protocol has best function in comparison with other protocol.

6. Protocol Analysis by CPN

We used a Petri net to model our security protocol. Colored Petri Nets (CP-nets or CPNs) is a graphical We used a Petri net to model our security protocol.
Colored Petri Nets (CP-nets or CPNs) is a graphical language for constructing models of concurrent systems

Table 1: Attack in authentication protocol.

attack	PKMv2(with nonce)	EAP	Proposed Protocol
MITM	About weak	resistant	resistant
Replay	About weak	resistant	resistant
Interception	resistant	resistant	resistant

and analysing their properties. CPnets is a discrete-event modeling language combining Petri nets and the functional programming language CPN ML which is based on Standard ML . A CPN model of a system describes the states of the system and the events (transitions) that can cause the system to change state. By making simulations of the CPN model, it is possible to investigate different scenarios and explore the behaviors of the system. Very often, the goal of simulation is to debug and investigate the system design. CP-nets can be simulated interactively or automatically[11]. Petri nets are composed from graphical symbols designating places (shown as circles), transitions (shown as rectangles), and directed arcs (shown as arrows). The Petri net model is illustrated in Figure 7. The model is simulated with the time color Petri net simulation tool. The basic information about the size of the state space and standard behavioral properties of the CPN model can be found in the state space report. For the CPN model of the proposed protocol in this study, the state space report is shown in Figure 6. As shown in Figure 6, we have data about "State Space statistics (Strongly-connected-component/Scc graph)", "Liveness Properties (Dead Markings, Dead Transition Instances, and Live Transition Instances)". The state space statistics inform about the size of the state space. For the model of proposed protocol, there are 11 nodes and 10 arcs. If the nodes and arcs in the state space and Scc graph are equal, it means that there are no cycles in the model. The number of nodes and arcs in the state space and Scc graph of two protocols are equal. It means that the token will not fall in a loop, and we have finite-occurrence sequences. A dead marking is a mark in which no element is enabled. This means that the marking corresponding to node 11 in Figure 7 is a dead marking. A transition is live if from any reachable marking we can always find an occurrence sequence containing the transition. As shown in Figure 6, there are no live transitions[10].Two protocols have a dead marking. So, they have not "live transitions". It is noted that no transition could be enabled from the dead marking. Also, there are no dead transitions in two protocols. A transition is dead, if there is no reachable marking in which it is enabled.There is no dead transition, which means that each transition in the protocol has the possibility to occur at least once. If a model has a dead transition, then it

```
Statistics
------------------

State Space
Nodes:   11
Arcs:    10
Secs:    0
Status: Full

Scc Graph
Nodes:   11
Arcs:    10
Secs:    0
```

Fig. 6 state space report.

corresponds to parts of the model that can never be activated. Hence, we can remove dead transitions from the model without changing the behavior of it. To compare the service delivery time of four protocols, we assume that each transition in two protocols takes 5 time units [10]. Several different kinds of output can be generated for data collector monitors. All of the data that is collected by a data collector can be saved in a data collector log file (Table.2).The log file also contains information about the steps and model times at which the data was collected [11].

Table 2: Timed static for propose protocol.

Name	Count	Avrg	90% Half Length	95% Half Length	99% Half Length	SSD	Variance	StD
Marking_size_simple'Data_Received_1	4	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Marking_size_simple'b_1	6	0.200000	0.332389	0.424105	0.665108	8.000000	0.163265	0.404061
Marking_size_simple'recive_ms_1	4	0.700000	0.760976	1.029080	1.889018	20.500000	0.418367	0.646813

Simulation steps executed: 10
Model time: 50.0

7. Conclusion

In this paper, we have introduced a authentication protocol for Wimax network. The proposed protocol provides mutual authentication, key exchange between MS and BS, so the probability of some threats such as eavesdropping, MITM and replay is reduced. Also, we have compared the performance of proposed protocol with another protocols by using CPN Tools. It has been shown that by omission of some transactions, the number of place and transition are reduced. Also, the execution time is reduced significantly, and the network resources are reserved. The state space reports are shown in Table 3.

Table 3: state space statistic.

property	Pkmv2(RSA)	EAP	proposed
Simulation step	7	13	10
Model time	40	90	50
Number of place	12	12	10
Number of transition	5	9	5
State space nodes	10	19	11
State space arc	9	22	10
Dead marking	20,21,22	6,19	11

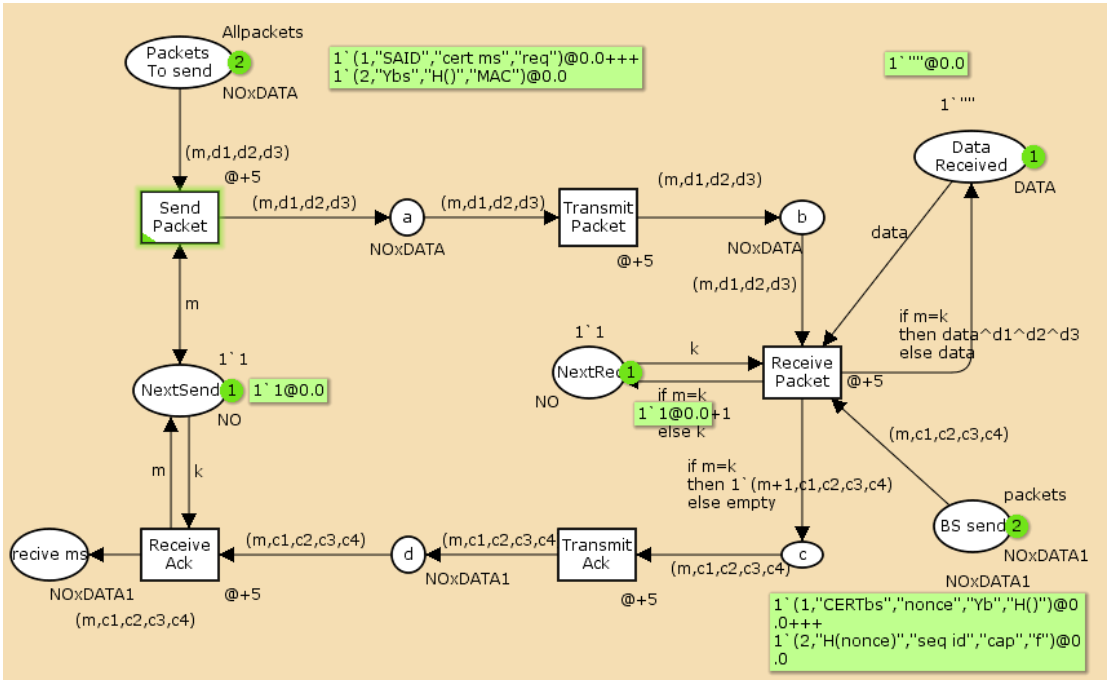


Fig. 7 petri net model for proposed protocol.

References

[1] E.Yuksel," Analysis of the PKMv2 Protocol in IEEE 802.16e-2005 Using Static Analysis", Technical University of Denmark,Feb 2007.

[2] N.Li," Research on Diffie-Hellman Key Exchange Protocol",IEEE,2010.

[3] L.Han,"a threat analysis of the Extensible Authentication Protocol",Apr 2006.

[4] J.Hur,H.Shim,P.Kim,H.Yun,N.Oak song,"security considerations for handover shemes in mobile wimax networks",IEEE,2008.

[5] A.Taha,A.Abdel hamid,A.Tahar,"formal verification of IEEE 802.16 security sublayer using scyther tools", ESRGroups France,2009

[6] Z.You,X.Xie,W.Zheng," Verification and Research of a Wimax authentication protocol Based on SSM",ICETC,2010.

[7] C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

[8] M. Bogdanoski, P.Latkoski, A.Risteski, B.Popovski," IEEE 802.16 Security Issues: A Survey",Telecommunication forum,2008.

[9] H.Tseng, R.Hong Jan, W.Yang,"A chaotic maps-base key agreement protocol that preseres user anonymity",IEEE ICC,2009.

[10] M.Shaikhan,A.Sobhani,M.E.Kalantari," Modification of Mobile Web Shopping Protocol Using GAA and Analysis by Colored Petri Nets",SATCCN,2011.

[11] K. Jensen , L.Michael Kristensen, L. Wells," Coloured Petri Nets and CPN Tools for Modelling and Validation of Concurrent Systems" ,Department of Computer Science,2008.

[12] J.Huang, C.Tser Huang," Secure Mutual Authentication Protocols for Mobile Multi-hop Relay WiMAX Networks against Rogue Base/Relay Stations"IEEE,2011.

[13] S.Sidharth,M.P.Sebastian," A Revised Secure Authentication Protocol for IEEE 802.16 (e)", International Conference on Advances in Computer Engineering,2010.
[14] M.Holbal,T,Welzer," An Improved Authentication Protocol Based on One-Way Hash Functions and Diffie-Hellman Key Exchange", International Conference on Availability, Reliability and Security,2009.
[15] F.Leu, Y. Huang, C.Hong Chiu," Improving security levels of IEEE802.16e authentication by Involving Diffie-Hellman PKDS", International Conference on Complex, Intelligent and Software Intensive Systems,2010.
[16] Ergang Liu, Kaizhi Huang and Liang Jin,"the design of trusted access scheme base on identity for wimax network" IEEE computer society (International Workshop on Education Technology and Computer Science),2009

Mohammad zabihi recived the B.S. in 1997 and M.S. degree in 2011 in communication engineering from Islamic azad university-shahre rey branch. Now he is an expert in telecommunication company. his research interests include network security, mobile communication and NGN networks.

Ramin Shaghaghi Kandovan received the B.S. degree in electronic engineering from Tehran University, Tehran, Iran, in 1990 and M.S. and Ph.D. degrees in communication engineering from Islamic Azad University, Tehran, Iran, in 1993 and 2002, respectively. He is currently an Associate Professor in Communication Engineering Department of Islamic Azad University-Shahre Rey Branch. His research interests include Higher Order Statistics, security in communication networks, signal processing. He has been the Head of Post-Graduate Center of IAU-Shahre Rey Branch since 2009.

Mohammad Esmail Kalantari received the B.S. degree in communication engineering from Communication Technical Faculty, Tehran, Iran in 1972 and M.S. and Ph.D. degrees in communication engineering from Ecole National Superieur des Telecommunications (ENST), Paris, France, in 1979 and 1982, respectively. He has been an Assistant Professor in Electrical Engineering Department of Khaje Nasir Toosi University of Technology for 30 years. Now, he is an academic member of Islamic Azad University, Shahre-Rey Branch, Iran. His research interests include security in communication networks, next generation networks, and mobile communication systems.

PPNOCS: Performance and Power Network on Chip Simulator based on SystemC

El Sayed M. Saad[1], Sameh A. Salem[1], Medhat H. Awadalla[1,2], and Ahmed M. Mostafa[1]

[1] Communication, Electronics and Computers Department, Faculty of Engineering, Helwan University, Helwan, Egypt

[2]Electrical and Computer Engineering Department, SQU University, Oman

Abstract

As technology moves towards multi-core system-on-chips (SoCs), networks-on-chip (NoCs) are emerging as the scalable fabric for interconnecting the cores. Network-on-Chip architectures have a wide variety of parameters that can be adapted to the designer's requirements. This paper proposes a performance and power network on chip simulator (PPNOCS) based on SystemC to explore the impact of various architectural level parameters of the on-chip interconnection network elements on its performance and power. PPNOCS supports an arbitrary size of mesh and torus topology, adopts five classic routing algorithms and seven synthetic traffic patterns. Developers also can develop and verify their own network design by modifying the corresponding modules. Experiments of using this simulator are carried out to study the power, latency and throughput of a 4x4 multi-core mesh network topology. Results show that PPNOCS provides a fast and convenient platform for researching and verification of NoC architectures and routing algorithms.

Keywords: *Network-on-Chip, Performance, Power, Simulation, SystemC.*

1. Introduction

Networks-on-chip [1] are critical elements of modern system-on-chip as well as multi-core designs. They consist of routers, links, and well-defined network interfaces. Packet-switched interconnection networks [2] facilitate communication between cores by routing packets between them. The structured and localized wiring of such a NoC design simplifies timing convergence and enables robust design that scales well with device performance.

One major difficulty that faces NoC architects is to select a communication network that suits a specific application or a range of specific applications with the constraints of cost, power and performance. Design decisions are typically made on the basis of simulation before resorting to emulation or implementation since it is cheap and flexible. To make a right decision on the network architecture, a simulation tool should enable to faster explore the architectural design space and assess design quality regarding performance, cost, and power.

SystemC [3] and Transaction Level Modeling (TLM) [4] have become quite popular and have found a relatively wide range of applications both in academia and industry [5]. SystemC is an extension of C++, in the form of a hardware-oriented library of C++ classes [6]. TLM is a library of functions built on the top of SystemC. In the TLM terminology, a transaction represents the information being exchanged between the different system modules. TLM is particularly interested in separating the computational component from the communication component. For this purpose, TLM provides constructs to efficiently model the inter-module communication such as channels, interfaces and ports, which are objects provided by SystemC.

This paper presents a performance and power network on chip simulator (PPNOCS) based on SystemC, to explore the impact of various architectural level parameters of the on-chip interconnection network elements on its performance and power. A general modularized NoC node structure is first realized under SystemC, and then connected to form the network. Users also can develop their own network topology and routing algorithm by modifying the corresponding modules. Then they can verify their design by loading different network traffic patterns to run the simulation.

The paper is organized as follows: Section 2 provides a brief overview of related work. The simulation platform is described in Section 3. Experimental results are discussed in Section 4. Finally, Section 5 concludes the paper.

2. Related Work

With the emergence of the NoC concept, researchers have realized the need to evaluate NoC systems. This has led to the use of existing network simulators, which have been adapted for on-chip communication networks [7]. Xu et al. employed the OPNET network simulator for simulation of on-chip network systems [8]. Such an approach leverages the already existing tool, which has had time to mature.

However, on-chip communication is different than traditional networks and parallel computer communication networks. NoC simulation environment must accurately reflect on-chip behaviors. Nostrum is another attempt of NoC simulation developed at KTH, Stockholm and it offers a packet switched communication platform based on the traditional OSI model of computer networks [9]. Initially, mesh topology is selected to prove the concept of Nostrum simulator. Recently, attempts have also been made to extend Nostrum to support both regular and irregular NoC topologies [10].

Many simulation tools have been developed to research the design of router architectures [11, 12] and NoC topologies [13] with varying area/performance [14] trade-offs for general purpose SoCs. Kogel et. al. [15] presents a modular exploration framework to capture performance of point-to-point, shared bus and crossbar topologies. The impact of varying topologies, link and router parameters on the overall throughput, area and power consumption of SoCs using relevant traffic models is discussed in [16]. Orion [17] is a power-performance interconnection network simulator that is capable of providing power and performance statistics. Orion model estimates power consumed by router elements (crossbars, FIFOs and arbiters) by calculating switching capacitances of individual circuit elements. Most of these tools do not allow for exploration of the various link level options of wire width, pitch, serialization, repeater sizing, pipelining, supply voltage and operating frequency.

In [18], Madsen et al. presented a NoC model which, together with a multiprocessor real-time operating system (RTOS) are used to model and analyze the behavior of a complex system that has a real-time application running on it. Mesh and torus are implemented in their design. Nurmi et al. [19] proposed a simulation environment by creating a library of pre-designed communication blocks that can be selected from a component library and configured by automated tools. From simulation point of view, these simulation tools are flexible to perform NoC design exploration. However, they are limited in topologies, and performance metrics [20].

In this paper, the proposed simulation platform is built from the ground up for Network-on-Chip simulation. The platform is built in SystemC, and takes advantage of the low-level modeling available in SystemC communication primitives, while leveraging the efficiency of C++ to achieve a balance between accuracy and performance. The main contributions of our simulation platform include the following:

- Explore the impact of various architectural level parameters of the on-chip interconnection network elements on its performance and power.
- Owing to the general NoC node structure and modularization modeling, users can extend the simulator with their own routing algorithm and network topology.
- PPNOCS provides a fast and convenient platform for researching and verification of various Network-on-Chip architectural designs.

3. Simulation Platform

A wormhole-router provides the necessary fine-grained flow control in terms of buffer and latency requirements, while the addition of virtual-channels aids in boosting performance and circumventing message-dependent deadlock [21]. Furthermore, Quality-of-Service (QoS) enhancements can be achieved by prioritizing the allocation of virtual-channels and switch bandwidth. For these reasons, PPNOCS implements the generic virtual-channel router shown in figure 1.

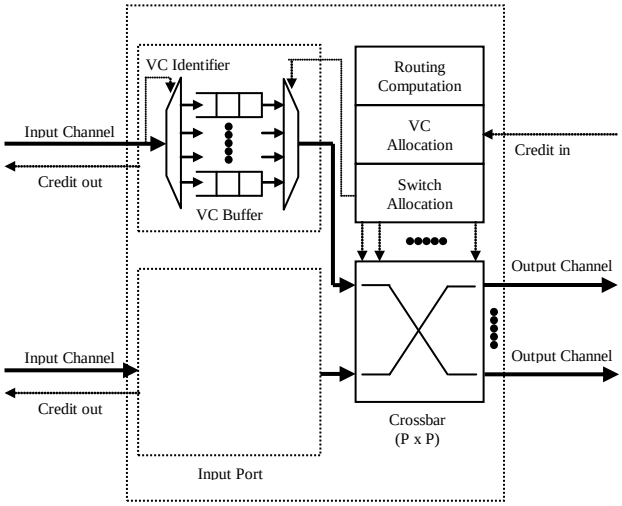


Fig. 1 Virtual-Channel Router

The router has P input ports and P output ports, supporting N virtual-channels (VCs) per port. Virtual-channel flow control exploits an array of buffers at each input port. By allocating different packets to each of these buffers, flits from multiple packets may be sent in an interleaved manner over a single physical channel. This improves both throughput and latency by allowing blocked packets to be bypassed.

3.1 PPNOCS Node Structure

To enable easy extensibility of the simulation platform, PPNOCS develop a modularized architecture for the generic router. Figure 2 shows the developed PPNOCS node structure.

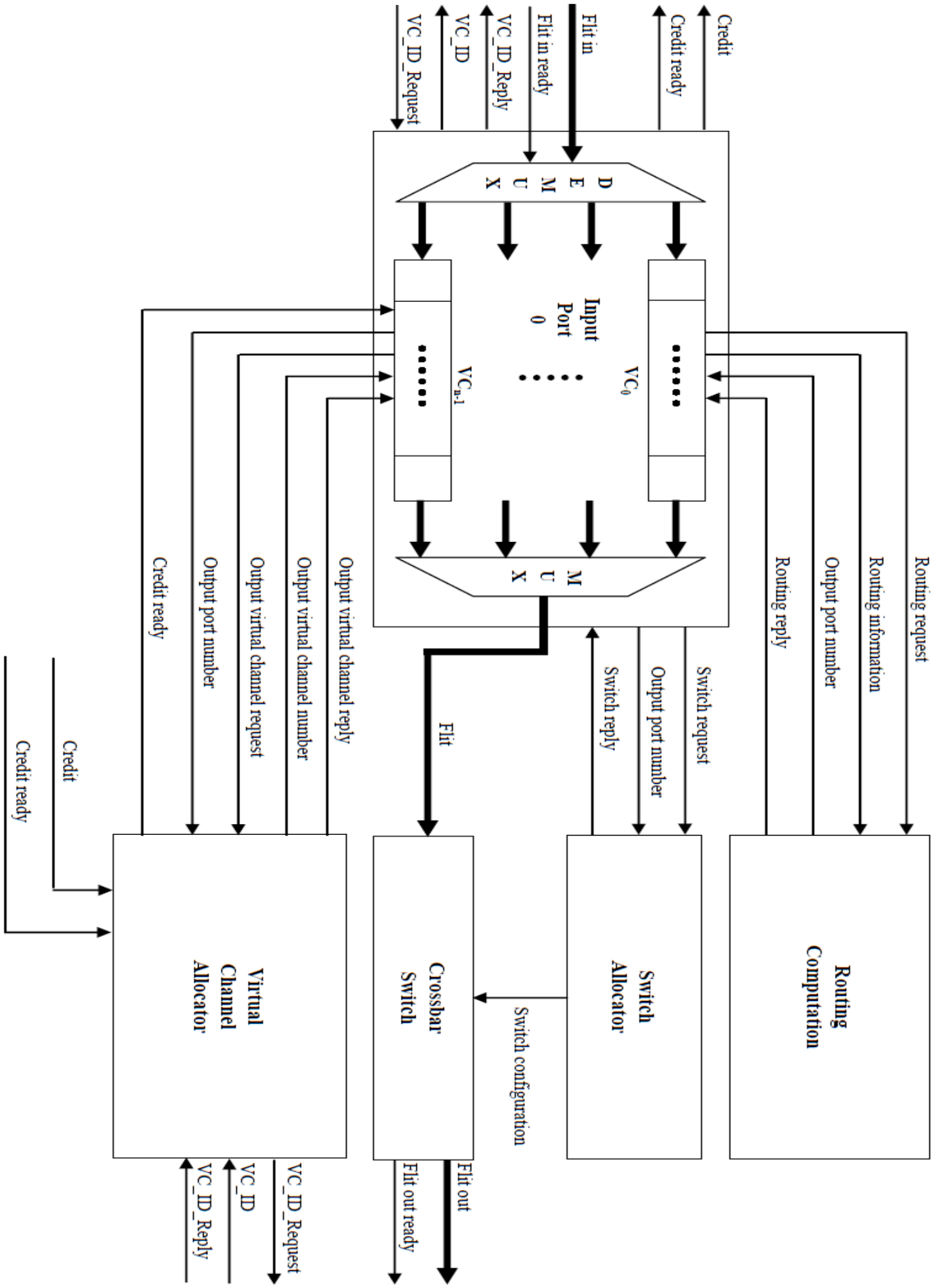


Fig. 2 PPNOCS node structure

In PPNOCS, a packet is divided into flow control digits or flits. A flit is the basic unit of bandwidth and storage allocation used by most flow control methods. The position of a flit in a packet determines whether it is a head flit, body flit, or tail flit. A head flit is the first flit of a packet and carries the packet's routing information. A head flit is followed by zero or more body flits and a tail flit. In a very short packet, there may be no body flits. In the following, a brief description of each module in the PPNOCS node structure is given.

3.1.1 Input Port Module

Input port module consists of a set of virtual channel modules. Each virtual channel consists of a FIFO buffer. The user can determine the number of virtual channels and the depth of each buffer in terms of flits. Any flit arrives at the input port contains a virtual channel identifier (VC_ID) which determines in which virtual channel buffer it will be stored.
All of the flow control mechanisms that use buffering need to know the availability of buffers at the downstream nodes. Then the upstream nodes will determine when a buffer is available to hold the next flit to be transmitted. This type of buffer management provides backpressure by informing the upstream nodes when they should stop flit transmission because all of the downstream flit buffers are full. Three types of low-level flow control mechanisms are in common use today to provide such backpressure: credit-based, on/off, and ack/nack [22]. PPNOCS implements the credit-based flow control mechanism. With credit-based flow control, the upstream router keeps a count of the number of free flit buffers in each virtual channel downstream. Then, each time the upstream router forwards a flit, thus consuming a downstream buffer, it decrements the appropriate count. If the count reaches zero, all of the downstream buffers are full and no further flits can be forwarded until a buffer becomes available. Once the downstream router forwards a flit and frees the associated buffer, it sends a credit to the upstream router for incrementing the buffer count.

3.1.2 Routing Computation Module

When a head flit of a new packet arrives at the input port, a routing request along with the routing information is sent to the routing computation module. According to the routing algorithm a set of valid output ports is produced and sent back to the input port module. The number of outputs produced by the routing computation module will depend on the routing algorithm. If more than one output produced, the selection function randomly select one of these outputs. PPNOCS implements five routing algorithms:

XY Routing Algorithm: XY routing is a dimension ordered routing which routes packets first in x- or horizontal direction to the correct column and then in y- or vertical direction to the receiver. XY routing suits well on a network using mesh or torus topology. Addresses of the routers are their xy-coordinates. XY routing never runs into deadlock or livelock [23]. Figure 3 shows an example of XY routing.

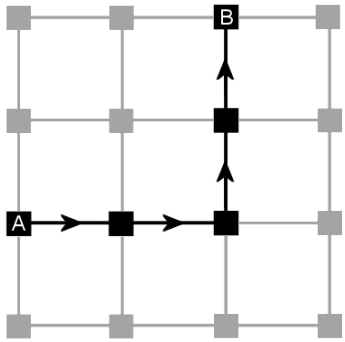


Fig. 3 XY routing from router A to router B

West-First Routing Algorithm: A west-first routing algorithm prevents all turns to west. So the packets going to west must be first transmitted as far to west as necessary. Figure 4 shows the allowed turns in the west-first routing.

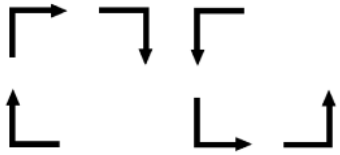


Fig. 4 Allowed turns in west-first routing

North-Last Routing Algorithm: Turns away from north are not possible in a north-last routing algorithm. Thus the packets which need to be routed to north must be transferred there at last. Figure 5 shows the allowed turns in the north-last routing.

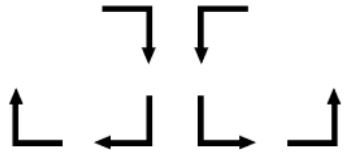


Fig. 5 Allowed turns in north-last routing

Negative-First Routing Algorithm: Negative-first routing algorithm allows all other turns except turns from positive direction to negative direction. Packet routings to negative directions must be done before anything else [24]. Figure 6 shows the allowed turns in the negative-first routing.

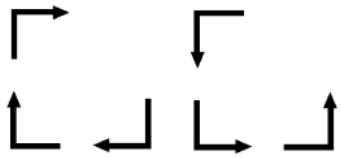


Fig. 6 Allowed turns in negative-first routing

Fully Adaptive Routing Algorithm: Fully adaptive routing algorithm uses always a route which is not congested. The algorithm does not care although the route is not the shortest path between sender and receiver [22].

3.1.3 Virtual Channel Allocation Module

After selecting a specific output port for the packet, the input port module sends a virtual channel request along with the output port number to the virtual channel allocation module. Then, the virtual channel allocator module sends a virtual channel request to the specified input port module in the downstream router. After receiving the VC_ID from the downstream router, the virtual channel allocator module sends it back to the input port module.

3.1.4 Switch Allocation Module

Each flit waiting in a virtual channel buffer and has available space in the downstream buffer can send a switch request to the switch allocator module. PPNOCS implements 5 × 5 (5-input × 5-output) input-first separable allocator. In an input first separable allocator, arbitration is first performed to select a single request at each input port. Then, the outputs of these input arbiters are input to a set of output arbiters to select a single request for each output port. The result is a legal matching, since there is at most one grant asserted for each input and for each output. The switch allocator module sends a switch reply for each input port module wins in the arbitration. If multiple VCs in the input port have been requested the same output port which is granted by the switch allocator, then they will be serviced in a Round Robin (RR) fashion. Upon granting the switch allocation requests, the switch allocation module sends a switch configuration signals to the crossbar switch module.

3.1.5 Crossbar Switch Module

Flits that have been granted passage on the crossbar are passed to the appropriate output ports.

3.2 Traffic Patterns

Application-driven workloads can be too cumbersome to develop and control [22]. This motivates the inclusion of synthetic workloads, which capture the salient aspects of the application-driven workloads, but can also be more easily designed and manipulated. Synthetic workloads are divided into three independent aspects: traffic patterns, injection processes, and packet length.

Traffic pattern is the spatial distribution of messages in interconnection networks. This message distribution is represented with a traffic matrix, where each matrix element $\lambda_{s,d}$ gives the fraction of traffic sent from node s destined to node d. Table 1 lists some common static traffic patterns used to evaluate interconnection networks [22].

Table 1: Network traffic patterns

Name	Pattern
Random	$\lambda_{sd} = 1/N$
Bit complement	$d_i = \neg s_i$
Bit reverse	$d_i = s_{b-i-1}$
Bit Rotation	$d_i = s_{i+1 \bmod b}$
Shuffle	$d_i = s_{i-1 \bmod b}$
Transpose	$d_i = s_{i+b/2 \bmod b}$

PPNOCS supports seven synthetic traffic patterns (Uniform Random, Hotspot, Bit Reversal, Bit Complement, Bit Rotation, Shuffle, and Matrix Transpose).

Random traffic: In which each source is equally likely to send to each destination is the most commonly used traffic pattern in network evaluation. Random traffic is very benign because, by making the traffic uniformly distributed, it balances load even for topologies and routing algorithms that normally have very poor load balance. Some very bad topologies and routing algorithms look very good when evaluated only with random traffic [22].

Hotspot Traffic: In hotspot traffic pattern, there's a particular node that will receive more traffic than other nodes. In PPNOCS, the hotspot is specified along with the percentage of the traffic dedicated to it.

Bit Reversal, Bit Complement, Bit Rotation, Shuffle, and Matrix Transpose: These are called Bit permutation patterns, in which each bit d_i of the b-bit destination address is a function of one bit of the source address, s_j where j is a function of i. permutation traffic patterns stresses the network topology or the routing algorithm because each source s sends all of its traffic to a single destination.

Injection process determines the average number of packets it injects per cycle (injection rate). The most common injection processes used in network simulations is the Bernoulli process [22]. For a Bernoulli process with rate r, the injection process A is a random variable with the probability of injection a packet equal to the process rate,

P(A = 1) = r. PPNOCS implements the Bernoulli process and the user can specify the packet injection rate before running the simulation. Also, the packet length in terms of flits can be specified.

3.3 Architecture Parameters

This section summarizes the different architectural parameters that can be configured before running the simulation. Table 2 shows a brief description for each parameter that can be specified by PPNOCS.

Table 2: Network Architectural Parameters

Parameter Name	Description
Topology	2D Mesh or 2D Torus
DimX	Number of columns
DimY	Number of rows
NUM_INPUTS	Number of input and output ports
VC_NUM	Number of virtual channels in each input port
VC_BUFFER_SIZE	Number of buffers for each virtual channel
PACKET_INJECTION_RATE	Injection rate (< =1)
TRAFFIC_DISTRIBUTION	Uniform Random, Hotspot, Bit Reversal, Bit Complement, Bit Rotation, Shuffle, and Matrix Transpose
ROUTING_ALGORITHM	XY, West-First, North-Last, Negative-First, and Fully Adaptive
PACKET_SIZE	Number of flits in the packet
WARM_UP_TIME	The number of clock cycles after which the simulator starts to collect statistics
SIMULATION_TIME	The number of clock cycles that have to be simulated.

3.3 Performance Metrics

A standard set of performance metrics can be used to compare and contrast different NoC architectures. The performance metrics evaluated by PPNOCS include throughput and packet latency [22].

3.3.1 Throughput

Throughput is the rate at which packets are delivered by the network for a particular traffic pattern. It is measured by counting the packets that arrive at destinations over a time interval for each flow (source-destination pair) in the traffic pattern and computing from these flow rates the fraction of the traffic pattern delivered [22]. Throughput, or accepted traffic, is to be contrasted with demand, or offered traffic, which is the rate at which packets are generated by the Intellectual Property (IP) block. Throughput can be defined as follows [16]:

$$\frac{\text{Total number of packets received at their destinations}}{\text{(Number of IP blocks) x (Total Time in Cycles)}}$$

3.3.2 Packet latency

Transport latency is defined as the time (in clock cycles) that elapses from between the occurrence of head flit injection into the network at the source node and the occurrence of the tail flit reception at the destination node [25].
In order to reach the destination node from some starting source node, flits must travel through a path consisting of a set of routers and interconnects [19]. Depending on the source/destination pair and the routing algorithm, each packet may have a different latency [19]. Therefore, for a given packet P_i, the latency L_i is defined as:

L_i = receiving time (tail flit of P_i) − sending time (head flit of P_i)

Let F be the total number of packets reaching their destination IPs and let L_i be the latency of packet P_i, where i ranges from 1 to F. The average packet latency, L_{avg}, is then calculated according to the following equation [19]:

$$L_{avg} = \frac{\sum_{i=1}^{F} L_i}{F}$$

3.4 Power Model

In PPNOCS, a power model at flit level is proposed depending on the results obtained from the Intel 80-core teraflop chip [26]. To explain the proposed model, consider a source IP injects a head flit into the write port of the input port module. The virtual channel module writes the flit into the tail of the FIFO buffer and emits a buffer write event, which triggers the buffer power model to compute buffer write power P_{write}. After the routing module determines the output port to which the head flit will be sent, a request is sent to the switch allocator module for the desired output port. The allocator module performs the required arbitration and generates an arbitration event, which signals the arbiter power model to compute arbitration power $P_{arbiter}$. Assuming the request is granted, the arbitration result is sent to the config port of the crossbar module. A grant signal is also sent to the grant port of the virtual channel module, leading to the read port of the buffer module activated. The flit is then read, emitting a buffer read event, which causes the buffer power model to compute buffer read power P_{read}. The flit next

traverses the crossbar, from input port to the output port. The crossbar module emits a crossbar traversal event and the crossbar power model computes traversal energy $P_{crossbar}$. Finally, the flit leaves the router and traverses the link. The link module emits a link traversal event, which calls the link power model to compute link traversal power P_{link}.

The total power consumed by this head flit at this node and its outgoing link is as follows:

$$P_{flit} = P_{write} + P_{arbiter} + P_{read} + P_{crossbar} + P_{link} \quad (1)$$

Let:

$$P_{buffer} = P_{write} + P_{read} \quad (2)$$

Then by substituting Eq. (2) in Eq. (1):

$$P_{flit} = P_{buffer} + P_{arbiter} + P_{crossbar} + P_{link} \quad (3)$$

The Intel 80-core teraflop chip recently introduced by Intel [26] is a good example of an aggressive NoC prototyping effort. The Teraflops Processor architecture contains 80 tiles arranged in a 8 x 10 2D array and connected by a mesh network that is designed to operate at 5 GHz. A tile consists of a processing engine connected to a five-port router, which forwards packets between tiles. The communication power is significant at 28% of each processing tile's total power. As shown in Figure 7, clocking power, 33%, is the largest component of router power, with the FIFO buffers the second largest component at 22%. Power due to physical links, crossbar switch, and arbiter come next at 17%, 15% and 7%, respectively.

According to the proposed power model, it is required to compare and contrast different NoC architectures in terms of power consumed in buffers, links, crossbar, and arbiter. So, in PPNOCS, it is considered that a unit power (U_p) is consumed by a flit to traverse from the input port to the output port of the router and leave through the outgoing link. U_p is then divided between buffering, arbitration, crossbar traversal, and link traversal according to the power ratios presented by the Intel 80-core teraflop chip and shown in Figure 7.

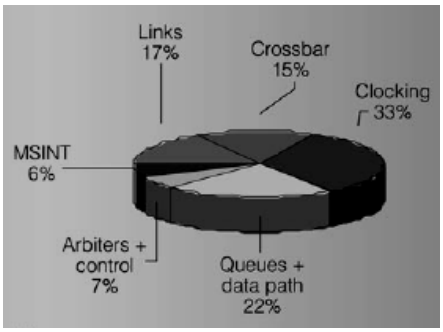


Fig. 7 Router power breakdown at 4 GHz, 1.2 V, and 110 C

4. Experimental Results

There are three potential ways of using PPNOCS for rapid exploration of network microarchitectures.

- The architect may wish to explore the impact of two application traffic patterns on specific network microarchitecture.
- The architect may wish to trade-off two configurations of microarchitecture, exploring their effect on network power and performance. This involves setting the network architectural parameters for the two configurations. Given a specific traffic pattern of the targeted application, the architect can feed the traffic pattern and configurations into two different instances of PPNOCS, and obtain their power and performance numbers.
- The architect may develop new network microarchitecture and wish to explore its impact on power and performance, evaluating it against a base microarchitecture. Owing to the general NoC node structure and modularization of PPNOCS, the architect can extend the simulator easily with their own microarchitectures.

In the experiments, a 16-node network organized as a 4 x 4 mesh is implemented, as shown in Figure 8. Each router has five physical bidirectional ports (north, south, east, west, and injection/ejection). Each simulation is run for a warm-up phase of 1000 cycles and simulation phase of 10000 cycles.

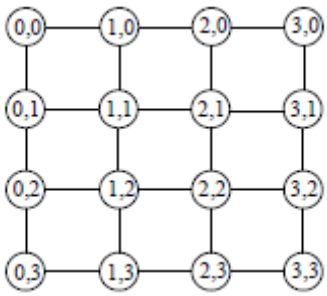


Fig. 8 A 4×4 2D mesh network

4.1 Exploring the effect of different traffic patterns

In this experiment, a 4×4 mesh NoC with XY routing is implemented, and 6 different traffic patterns to run the simulation is loaded. The packet length is two flits and each input port has 4 virtual channels with FIFO buffer

depth equal to 4 packets. The result of this simulation is shown in Figure 9.

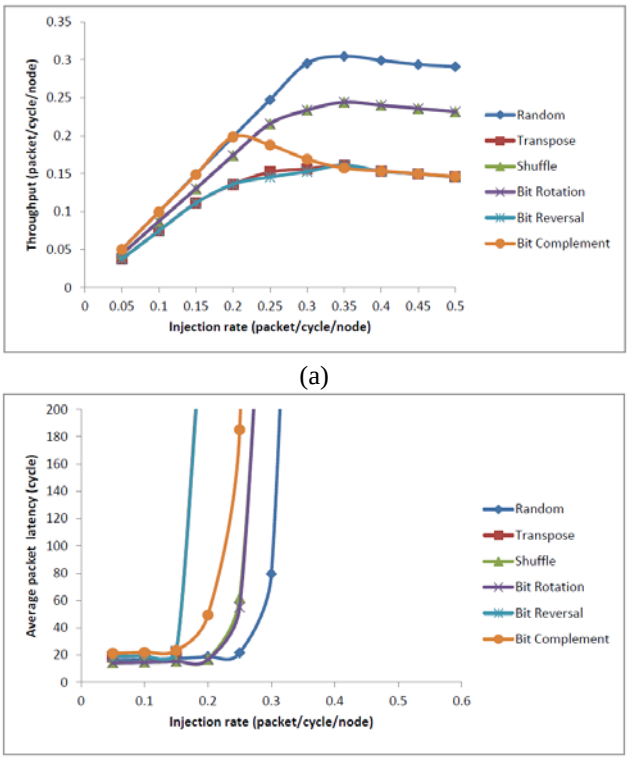


Fig. 9 Results of different traffic patterns

As shown in figure 9, Random traffic pattern gives better throughput and average packet latency according to its uniform and balanced distribution of load.

4.2 Exploring the effect of different configurations

This set of simulations is based on 4X4 mesh with XY routing, packet size equal two flits and under uniform random traffic. In this experiment, four different router configurations are simulated and compared:-

- Wormhole router with 64-flit input buffer per port (WH64).
- Virtual-channel (VC) router with 2 VCs per port and 16-flit input buffer per VC (VC_2_16).
- Virtual-channel router with 4 VCs per port and 4-flit input buffer per VC (VC_4_4).
- Virtual-channel router with 8 VCs per port and 8-flit input buffer per VC (VC8_8).

Figure 10 shows results obtained from simulating these routers in PPNOCS. Figure 5(a) shows VC_8_8 out-performing WH64, despite having the same total buffer size per input port, saturating at a higher packet injection rate of 0.35 packets/cycle/node. However, this performance improvement is achieved at the expense of higher power consumption, as indicated by Figure 11.

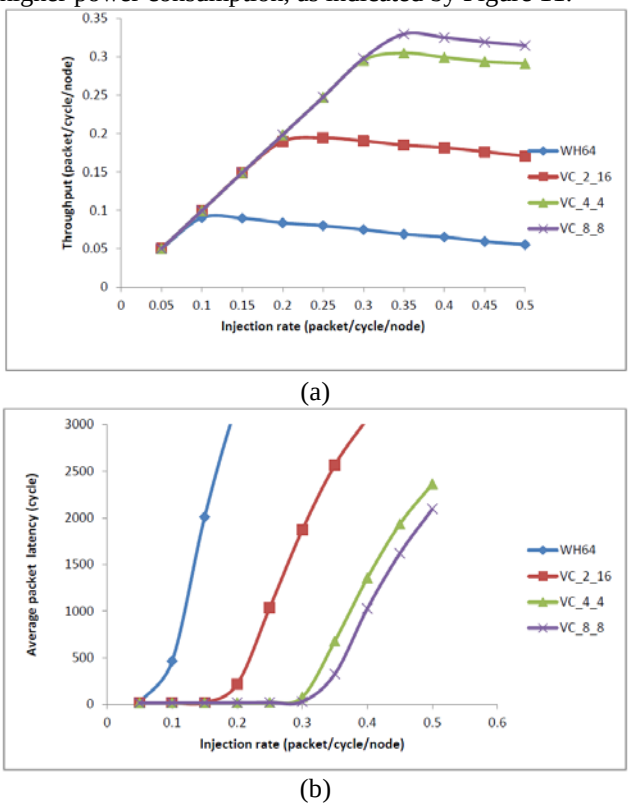


Fig. 10 Results of different configurations

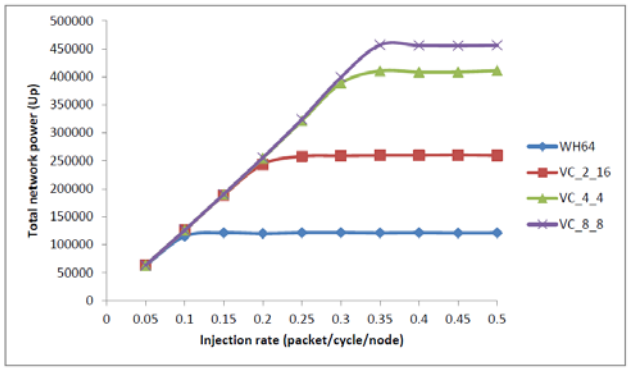


Fig. 11 Results of different configurations on power consumption

Beyond packet injection rate of 0.1 packets/cycle/node, VC_8_8 starts to consume more power than WH64, since it is still able to absorb the higher packet injection rate, so network activity continues to increase. For all configurations, total network power levels off after saturation, since the network cannot handle a higher packet injection rate, so the switching activity of the network remains constant.

It is interesting to note that VC_8_8 dissipates approximately the same amount of power as WH64 before saturation. Intuitively, since virtual-channel flow control is a more complicated protocol, requiring more complex hardware, we would expect a virtual-channel router to be more of a power hog than a wormhole router. The interpretation of this is that the power consumed by buffers, links and the crossbar switch is the dominant power consumption in a network node.
Figure 12 shows the power consumed by each component of the router. From these results, it can be verified that the power consumption of the buffer and crossbar components of the router is much more than the power consumed in arbitration.

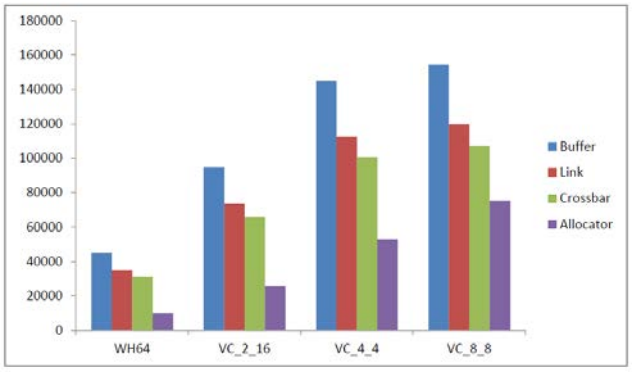


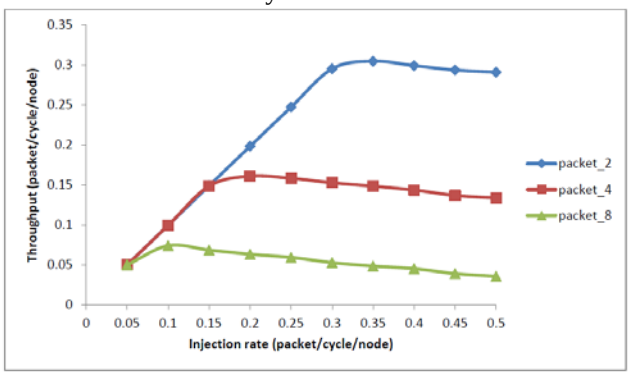
Fig. 12 Power consumption of different router components

4.3 Exploring the effect of different packet length

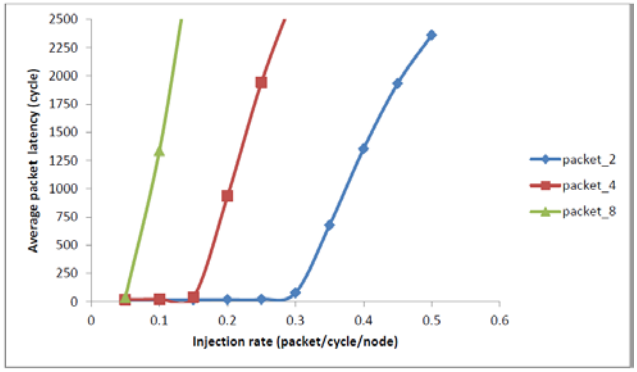
This set of simulations is based on 4X4 mesh with XY routing under uniform random traffic. Figure 13 shows the throughput and average packet latency for packet length equal to 2, 4, and 8 flits.
The throughput increase linearly when the injection rate is low. However, with the injection rate increasing, the confliction encountered in the network limits the increase of the throughput. Figure 13 shows that the average packet latency for 8 flits/packet is larger than that for 2 flits/packet. This is due to two reasons. First, longer packets will take more time to receive. Second, longer packets will cause more confliction at intermediate routers on the path from the source to the destination.

4.4 Exploring the effect of different routing algorithms

This set of simulations is based on 4X4 mesh with packet size equal two flits and under uniform random traffic. Table 3 lists the total number of received packets for the XY, West-First, North-Last, Negative-First, and Fully Adaptive routing algorithms for injection rates of 0.1 packets/cycle/node (under saturation), 0.3 packets/cycle/node (saturation), and 0.5 packets/cycle/node (above saturation). These results collected for the 10000 cycles simulation time.



(a)



(b)

Fig. 13 Results of different packet length

Table 3: Total number of received packets

Routing Algorithm	Injection Rate		
	0.1	0.3	0.5
XY Routing	15960	47226	46543
West-First Routing	16051	43662	42310
North-Last Routing	15992	44040	41915
Negative-First Routing	15933	43042	40653
Fully Adaptive Routing	15996	43546	41131

From the above table it can be seen that deterministic XY routing is faster than the other three partially adaptive algorithms. Partially adaptive algorithms can potentially speed up the time to deliver individual packets, but globally the results point out to poorer performance than the XY algorithm. Glass and Ni [27] suggested that reducing the number of turns that a message takes may reduce blocking and hence improve performance. This can be justified because adaptive routing has a trend to

concentrate the traffic at the center of the network, increasing in this way the number of blocked paths.

5. Conclusions

In this paper, a performance and power network on chip simulator (PPNOCS) based on SystemC has been proposed. As demonstrated, PPNOCS is a general NoC simulation and verification platform with high extensibility. Using PPNOCS, the impact of various architectural level parameters of the on-chip interconnection network elements on its performance and power can be explored. Owing to the general NoC node structure and modularization modeling, developers can develop their own routing algorithm and network topology in such a way that they can use either traffic patterns provided by PPNOCS or their own traffic pattern. Then, the simulation and design verification can be applied. By going through simulation experiments using five classic routing algorithms, the practical usage of PPNOCS is verified. Also, the impact of different traffic patterns, routing algorithms, virtual channel configurations, and packet length on the network performance and power is evaluated. As shown, PPNOCS provided a fast and convenient platform for researching and verification of NoC architecture and routing algorithm.

References

[1] L. Benini et al., "Networks on Chips: A New SoC Paradigm", Computer, vol. 35, no. 1, Jan. 2002, pp. 70-78.
[2] W.J. Dally and B. Towles, "Route Packets, Not Wires: On-Chip Interconnection Networks", Proc. 38th Design Automation Conf., ACM Press, 2001, pp. 681-689.
[3] http://www.systemc.org/, “Open systemc initiative.”
[4] Cai L, Gajski D, "Transaction-level modeling in system level design", CECS technical report (03-10), Center for Embedded Computer Systems, Information and Computer Science, University of California, Irvine, March 2003.
[5] Sandro Penolazzi, "A System-Level Framework for Energy and Performance Estimation of System-on-Chip Architectures", Ph.D. thesis, KTH School of Information and Communication Technology, Stockholm, Sweden, 2011.
[6] Thorsten Grotker, System Design with SystemC, Kluwer Academic Publishers, Norwell, MA, USA, 2002.
[7] Gul N. Khan and V. Dumitriu, "A Modelling tool for simulating and design of on-chip network systems", Embedded Hardware Design-Microprocessors and Microsystems, Vol. 34, No. 3-4, pp. 84-95, 2010.
[8] J. Xu, W. Wolf, J. Henkel, S. Chakradhar, "A design methodology for application specific networks-on-chip", ACM Transactions on Embedded Computing Systems, 2006, pp. 263–280.
[9] M. Millberg, E. Nilsson, R. Thid, S. Kumar, A. Jantsch, "The nostrum backbone – a communication protocol stack for networks on chip", in: Proceedings of the 17th International Conference on VLSI Design, 2004, pp. 693–696.
[10] L. Papadopoulos, S. Mamagkakis, S. Catthoor, D. Soudris, "Application – specific NoC platform design based on system level optimization", in: IEEE Computer Society Annual Symposium on VLSI, March 2007, pp. 311–316.
[11] K. Lee, S.-J. Lee, and H.-J. Yoo, "Low-power network-on-chip for high-perforamance soc design", IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 14, pp. 148–160, Feb. 2006.
[12] S. E. Lee, J. H. Bahn, and N. Bagherzadeh, "Design of a feasible on-chip interconnection network for a chip multiprocessor (cmp)", in Proc. Of, Computer Architecture and High Performance Computing. Intl. Symp. on, pp. 211–218, 2007.
[13] F. Karim et. al., "An interconnect architecture for networking systems on chips", IEEE Micro, vol. 22, pp. 36–45, Oct. 2002.
[14] Rehan Maroofi, Vilas Nitnaware, and Shyam Limaye, "Area Efficient Design of Routing Node for Network-on-Chip", International Journal of Computer Science Issues (IJCSI), Vol. 8, Issue 4, No 1, July 2011.
[15] T. Kogel et. al., "A modular simulation framework for architectural exploration of on-chip interconnection networks", in Proc. of, Hardware/Software Codesign and System Synthesis, 2003, pp. 338–351.
[16] P. P. Pande, C. Grecu, M. Jones, A. Ivanov, and R. Saleh, "Performance evaluation and design trade-offs for network-on-chip interconnect architectures", IEEE Transactions on Computers, vol. 54, pp. 1025–1040, Aug. 2005.
[17] H.-S. Wang, X. Zhu, L.-S. Peh, and S. Malik, "Orion: A power performance simulator for interconnection networks", in Proc. of, MICRO 35, 2002.
[18] J. Madsen, S. Mahadevan, K. Virk, and M. Gonzalez, "Network-on-chip modeling for system-level multiprocessor simulation", Proc. 24th IEEE Real-Time Systems Symp. (RTSS), 2003, pp. 265-274.
[19] D. Siguenza-Tortosa and J. Nurmi, "VHDL-based simulation environment for proteo NoC", Proc. 7th IEEE Int'l High-Level Design Validation and Test Workshop, 2002, pp. 1-6.
[20] Xinan Zhou, "Performance evaluation of network-on-chip interconnect architectures", M.S. thesis, Department of Electrical and Computer Engineering, University of Nevada, Las Vegas, 2009.
[21] Robert Mullins, Andrew West, and Simon Moore, "Low-latency virtual-channel routers for on-chip networks". In Proceedings of the International Symposium on Computer Architecture, 2004.
[22] W.J. Dally, B. Towles, "Principles and Practices of Interconnection Networks", Morgan Kaufmann, 2004.
[23] M. Dehyadgari, M. Nickray, A. Afzali-kusha, Z. Navabi, "Evaluation of Pseudo Adaptive XY Routing Using an Object Oriented Model for NOC", The 17th International Conference on Microelectronics, December 2005.
[24] H. Kariniemi, J. Nurmi, "Arbitration and Routing Schemes for On-chip Packet Networks", Interconnect-Centric Design for Advanced SoC and NoC, Kluwer Academic Publishers, 2004, pp. 253–282.
[25] P. P. Pande, C. Grecu, A. Ivanov, and R. Saleh, "Design of a switch for network on chip applications", Proc. Int'l Symp. Circuits and Systems (ISCAS), 2003, pp. 217-220.

[26] S. Vangal et al., "An 80-tile 1.28TFLOPS network-on-chip in 65 nm CMOS", in Proc. Solid-State Circuits Conf., Feb. 2007, pp. 98–589.
[27] Glass, C.; Ni, L., "The Turn Model for Adaptive Routing. Journal of the Association for Computing Machinery", v. 41(5), Sep. 1994, pp. 874-902.

RDWSN: To offer Reliable Algorithm for routing in wireless Sensor network

Arash Ghorbannia Delavar, Tayebeh Backtash and Leila Goodarzi

Department of Computer, Payame Noor Universtiy, PO BOX 19395-3697, Tehran, IRAN

Abstract

we would offer a reliable algorithm for routing in wireless sensor network (WSN). RDWSN algorithm by using of outstanding parameters has been evaluated compared with previous algorithms. The proposed algorithm with respect to size and distance in basic capability function has created a new target function that bears more effective than previous models. By use of RDWSN algorithm, we could improve wasted energy in WSN and also we could balance workload among CH with different paths. Parameters applied in the proposed algorithm may promote the reliability in order to make more balance. And in several simulations done by the same processor, the delay has been reduced with regard to previous algorithm in order to increase the reliability. Using a threshold detector with a combination of various parameters examined were able to help reduce delay in the area of sensor than previous algorithms and also improved power consumption and finally, The index parameters that were used in the capability function is increases sensors networks lifetime.

Keywords: *Delay, QOS , RDWSN, Reliability, routing ,Wireless Sensor Network.*

1. Introduction

Wireless sensor networks have been taken into consideration in recent years and used in different fields including medical care, industrial production, military applications and etc. Nodes are randomly distributed by uncontrollable devices in considered area and they form an adhoc network. It is natural that such a network may have hundreds or thousands of nodes. And nodes have some limitations including limited energy, low memory, limited computing power and none rechargeable batteries. Limitation of their bandwidth and short radio range. Managing a large number of nodes with these limitations can provide many challenges [2]. So, routing protocols should be in such a way that increase the lifetime of the network and improve service quality. Routing protocols are hierarchical and data collections imply to an organization based on cluster of sensors. In hierarchical structure, each cluster has a cluster head called CH [14]. And also, many of them are ordinary nodes (sn). Nodes of sensor transmit periodically their data to CH and CH transmits data to BS Station. Transmitting to BS can be done directly or through other CH. Whenever data is transmitted to the longer distance, energy consumption increases [9].
Regarding the mentioned cases, a lot of routing methods were created for WSNs which can be divided into three group's base on the most common categorization: Data–centric Algorithms, Location Base Algorithms and Hierarchical Algorithms .Data-centric protocols are query-based and depend on the naming of desired data, which
Helps in eliminating many redundant transmissions. Location-based Algorithms utilize the position information to relay the data to the desired regions rather than the whole network. Hierarchical Algorithms aim at clustering the nodes so that cluster heads can do some aggregation and reduction of data in order to save energy.
Cluster based methods benefit from a few characteristics: the first characteristic is that they divide the network into several parts and every part is directed by one cluster head. This characteristic causes the cluster based methods to be of a higher scalability level. The second characteristic is that a cluster head receives the data of its nodes and sends it BS after gathering data, which results in substantial reduction in data redundancy. We will provide a clustering algorithm, which uses a new distributed method, and a local threshold detector to perform clustering [1]. RDWSN algorithm performs routing among CH and by doing special techniques, it not only increases network life time but improves the reliability and end to end delay.

2. Related work

One of routing algorithms is Qosnet algorithm; in this algorithm for routing, a node bearing more energy is selected among other nodes.(Maxbv)and if such node dose not find among all nodes, the packet is discarded, therefore, the reliability decreases , and also end to end delay increases in order to find a node with more energy [3]. Another routing algorithm is MCMP algorithm, this algorithm has only considered to QOS and consumed energy by nodes being very important in wireless sensor networks and has not considered to the reliability and delay [4]. Another routing algorithm is DARA which uses the frequent packets for increasing the reliability, so this method causes to consume more energy in nodes for routing of frequent packets [7]. Also, algorithm MPDT has been offered to less energy consumption and life span of network that it does not consider to quality of services parameters. Many algorithms regarding the deduction of energy consumption have been offered in wireless sensor networks, such as:

SR: this protocol selects the paths based on strength and power of signals among nodes. Therefore, the paths selected are relatively stronger. This protocol can be divided into two parts. DRP is responsible for preparing and maintaining of routing table and the table related to the power of signals. SRP also studies received packets and if they have their own address, they would be transmitted to the higher layers [10,11]. Dynamic Source Routing (DSR): in this type, mobile nodes should provide temporary memories for the paths that are aware of their existence. Two main phases for this protocol has been considered: discovery of the path and updating path. [6]. Discovery of the path phase uses packet path request/reply and updating path phase uses confirmations and link mistakes. The temporarily ordered routing Algorithm (TORA)is based on distributed routing algorithm and has been designed for dynamic mobile networks. This algorithm for each pair of nodes will

determine several paths and require clock sync. Three main factors of this protocol are including: making path, updating path and destroying path. Ad hoc on demand distance vector (AODV) is based on algorithm DSDV; however, it would reduce emission because of routing at the time of necessity. Discovery of the path algorithm only starts its performance when a path does not exist between two nodes [8].

RDMAR: this type of protocol calculates the distance between two nodes through radio loops and navigation algorithms. This protocol determines the limited range of path, thus it reduces heavy traffic in network which it has less effective in improving quality of service and energy consuming [12, 13]. In Qosnet algorithm we had following equations:

$$B_{s_0 s_d}(Q) = \sum_{p \in Q} \sum_{s_i \in p} b_i(t) \quad (1)$$

$b_i(t)$ indicates Sensor battery S_i at the time of t. $B_{s_0 s_d}(Q)$ is the Battery Cost between source node S_0 and target node S_d

$$D_{s_0 s_d}(Q) = \min \{\sum d(s_i, s_{i+1})\} \quad (2)$$

$d(s_i, s_{i+1})$ indicates delay between sensor s_i and its neighbor s_{i+1} and end to end delay between S_0 and S_d is as follows:
According to Q is the collection of paths [3,4]
The reliability includes: the number of the received nodes in the target and the number of produced nodes in the source.
The reliability of end to end multi paths between S_0 and S_d in the collection of paths Q is as follows:

$$R_{s_0 s_d}(Q) = 1 - \prod_{p \in Q}(1 - r(p)) \quad (3)$$

r(p) the reliability of path is p.
D_{req} indicates end to end delay. R_{req} indicates the level of reliability Because we have (Q) $maxB_{s_0 s_d}$, must:

$$D_{s_0 s_d} \leq D_{req} \quad 4(a)$$
$$R_{s_0 s_d} \geq R_{req} \quad 4(b)$$

If we indicate the delay with L_i^d and the reliability with L_i^r in each link and h_i represents Hops and L_i^b is battery cost of each node, indicating:

$$L_i^d = \frac{D_{req} - D_i}{h_i} \quad 5(a)$$
$$L_i^r = \sqrt[h_i]{R_i} \quad 5(b)$$
$$L_i^b = \sum_{j \in f_w(s_i)} b_j(t) \quad 5(c)$$

D_i the experienced real delay in node S_i from the source node.
R_i is the requirements of the reliability given to the path via S_i.
$f_w(s_i)$ is the collection of pathfinder nodes[1].

2.1. RDWSN algorithm

In RDWSN algorithm, by using of relation which we will be explained in the following, we create a new target function and finally draw the proposed flowchart:

$$W_i = energy_i - b_i \quad (6)$$

W_i is the volume of a node.
$energy_i$ is the initial energy of a node.
b_i is the remaining energy of a node.
TD or threshold including the average of the highest energy and the lowest energy.

$$7)\ TD = \frac{maxE + minE}{2} \quad (7)$$

And finally, target function is offered by using of equations 5a, 5b, 6 as follows:

$$F = C_1 \frac{(D_{req} - D_i)}{h_i} + C_2 \sqrt[h_i]{R_i} + C_3 W_i + C_4 (dist_{ij})^L \quad (8)$$

3. proposal flowchart

Considering equation Nos. 7 and 8 , a flowchart is suggested as follows:
F_w (i) includes the collection of nodes that packet for reaching to Sink can pass from nodes; in the beginning, this collection is considered as empty set, because it has not been found a suitable node being in the ideal condition. D_{req} is the maximum delay which is tolerated to be reached packet from the source to Sink.
For this purpose, we add D_i which is equal to the delay in reaching packet from the source to node I, with d_{ij} (packet delay from node i to node), until the delay is calculated in reaching of packet from the source node to selected node i . and we consider collection of C_w(i) equal to the nodes that the delay in reaching packet to them is less than the requested amount D_{req} and residual energy of nodes must be more than threshold. if there is no node in this condition, network information would be updated; but if some nodes were accepted as candidate, we would select the most desirable node that is the nearest node to (minimum d_j) Sink with maximum reliability and minimum energy consumption and transmit the packet to such node. R_{req} is the minimum reliability to a path that packet covers the distance from the source node to Sink. The algorithm can be repeated up to Th.

To calculate the reliability, the path has used from equation $L^r_j = \sqrt[hj]{Rj}$ .And if this reliability is more than the requested reliability R_{req} , the node j considers as selected node. Also it should be noted to $\sum x_j \log(1 - R_{ij}) \leq \log(1 - L_i^r)$[4]. As it was noted, the energy of this selected node must be more than the amount of TD. This amount of threshold is calculated from equation7.

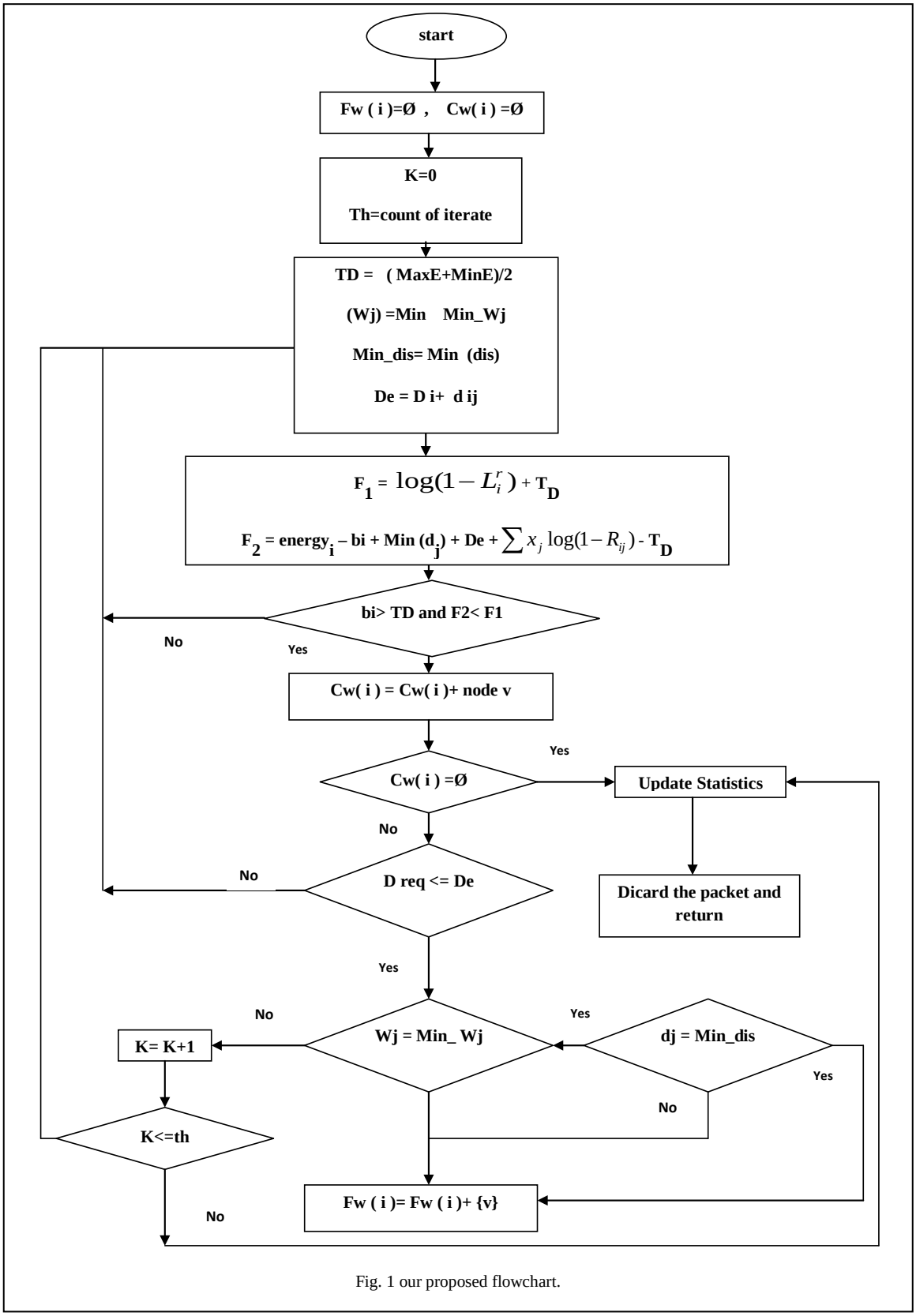


Fig. 1 our proposed flowchart.

4. Simulation

Simulation of RDWSN algorithm is performed in a restricted area of 200m *200m with about 200 nodes, We will analyze the presented algorithm in MATLAB, other simulation parameters has been displayed in table 1-1.

Table 1. Simulation of parameters

parameter	***value***
network area	200m*200m
sensors	70-200
transmission range	40m
bandwidth	250kbps
packet size	128b
simulation time	1000s
reliability	0-1
reporting rate	1 packet/s
delay	40-120

Also, to perform simulation, we need a system with hardware and software equipments that you can see this particulars in table No. 2. The nodes randomly distribute in this restricted area and Bs is installed 100 meter away from this certain area. The initial energy for all nodes has been considered 3.1 Joule. Figure 1 shows packet delivery ratio for 4 set algorithms.

Table 2.Minimum Hardware And Software Equipments

hardware or software	***Information***
Processor	Intel 1.8 GHz
Memory(RAM)	512 MB
operating system	Microsoft Windows XP
System type	32-bit Operating System

So the obtained reliability by RDWSN algorithm has improved in comparison with two algorithms MCMP and QOSNET. Of course, GODROUTING algorithm is our ideal algorithm in wireless sensor networks which has not been created up to now. Figure 2 indicates Average delivery ratio and as shown in diagram, reliability requirement has been very close RDWSN algorithm to ideal algorithm [3,4].
Figure 3 shows an average end to end delay for all algorithms in the same condition. As cleared in diagram, RDWSN algorithm after GODROUTING has the least delay among other algorithms. Figure 4 compares average delivery ration and delay requirement and in this comparison, RDWSN algorithm is better to two other algorithms.

Table 3. Simulation values

Number of nodes	***Average delivery ratio MCMP***	***Average delivery ratio QoSnet***	***Average delivery ratio RDWSN***	***Average delivery ratio GodRouting***
75	0.33	0.43	0.53	1
100	0.57	0.71	0.8	1
125	0.75	0.89	0.9	1
150	0.75	0.94	0.94	1
175	0.78	0.94	0.95	1
200	0.91	0.96	0.98	1
Reliability requirement	***Average delivery ratio MCMP***	***Average delivery ratio QoSnet***	***Average delivery ratio RDWSN***	***Average delivery ratio GodRouting***
0.7	0.9	0.97	0.98	0.99
0.75	0.9	0.97	0.98	0.99
0.8	0.9	0.97	0.98	0.99
0.85	0.91	0.97	0.99	0.99
0.9	0.93	0.99	1	1
0.95	0.95	0.99	1	1
Reliability requirement	***Average packet delay MCMP***	***Average packet delay QoSnet***	***Average packet delay RDWSN***	***Average packet delay GodRouting***
0.7	110	85	57	44
0.75	110	85	56	43
0.8	110	85	56	42
0.85	110	83	55	41
0.9	109	82	52	41
0.95	109	81	49	41
delay requirement	**Average delivery ratio MCMP**	**Average delivery ratio QoSnet**	**Average delivery ratio RDWSN**	**Average delivery ratio GodRouting**
60	0.28	0.55	0.61	1
70	0.49	0.78	0.85	1
80	0.7	0.88	0.89	1
90	0.82	0.92	0.93	1
100	0.88	0.94	0.94	1
120	0.91	0.98	0.99	1

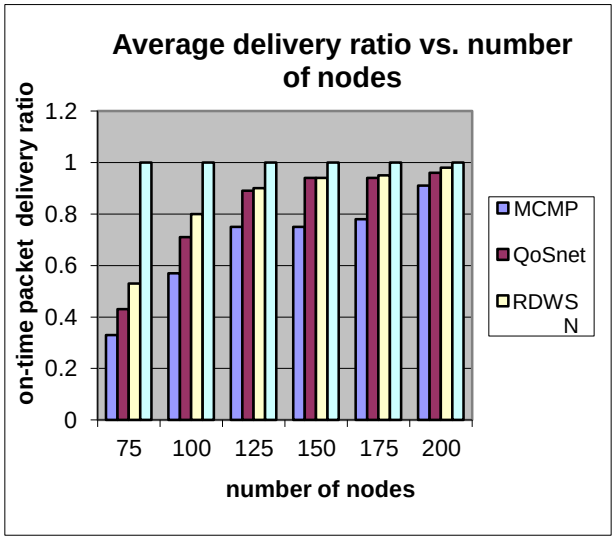


Fig. 2 Average delivery ratio VS. number of nodes.

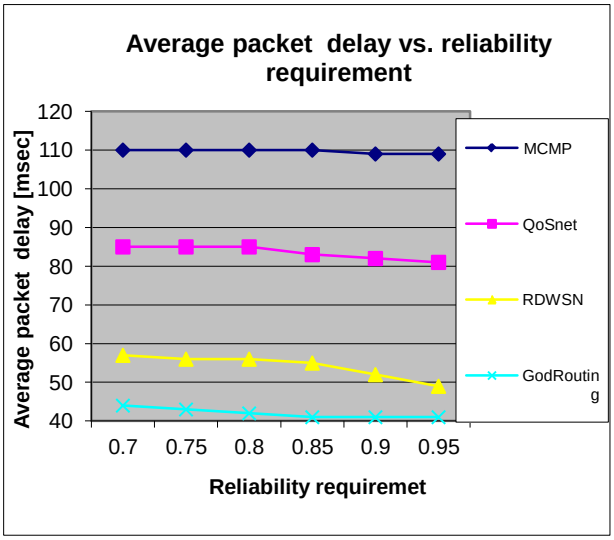


Fig. 4 Average packet delay VS. reliability requirement..

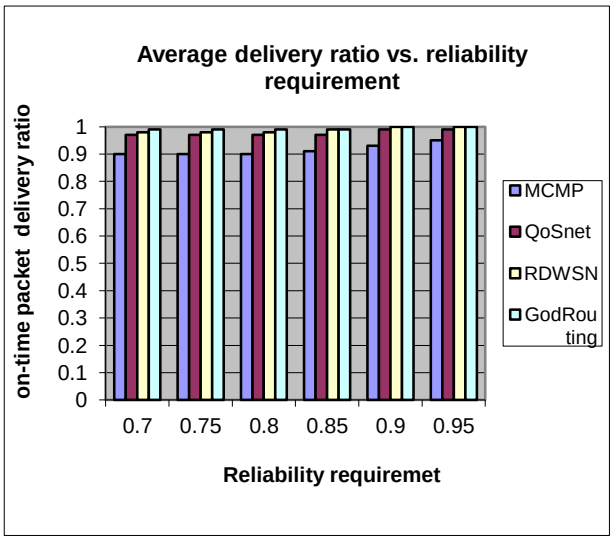


Fig 3. Average delivery ratio VS. reliability requirement.

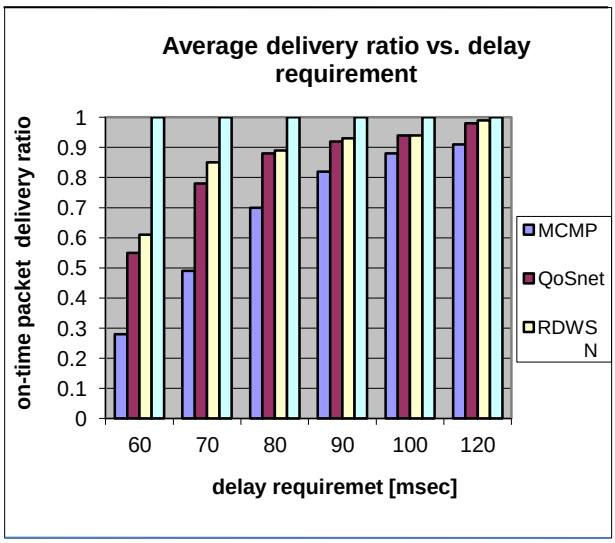


Fig 5. Average delivery ratio VS. delay requirement.

5. Conclusion

In this paper, we suggest the reliable algorithm for routing in sensor networks. RDWSN algorithm provides a reliable transmission environment with low energy consumption and less average end to end delay for transmitting packet toward BS. In this approach, TD is used for selecting next node that it is caused nodes with average energy to participate in the routing and also by weighting to nodes, we could improve the reliability and delay compared with two MSMP and QOSNET algorithms, for example, in the diagram of figure 2, as the evidence shows, by having number of different nodes, RDWSN algorithm at the best condition and special situation has increased the reliability of MCMP and QOSNET algorithms more than 0.23% and 0.10% respectively, and at the worst condition, it has been equal to QOSNET algorithm and in comparison with MCMP algorithm, it has improved at the rate of 0.07%. also considering diagram of figure 4, average delay of packets has decreased, so that with different reliabilities, average delay RDWSN at the best condition is less than MCMP and QOSNET algorithms .

We try to improve this algorithm in respect of other QOS parameters and approach to GODROUTING algorithm.

References

[1] Arash Ghorbannia Delavar, Javad Artin, and Mohammad Mahdi Tajari," PRWSN: A Hybrid Routing Algorithm with Special Parameters in Wireless Sensor Network", A. Özcan, J. Zizka, and D. Nagamalai (Eds.): WiMo/CoNeCo 2011, CCIS 162, pp. 145–158, 2011. © Springer-Verlag Berlin Heidelberg (2011)

[2] Arash Ghorbannia Delavar, Javad Artin, and Mohammad Mahdi Tajari," RCSDN : a Distributed Balanced Routing Algorithm with Optimized Cluster Distribution", 3rd International Conference on Signal Acquisition and Processing (ICSAP 2011) V2-368-V2-372

[3] Therence Houngbadji, Samuel Pierre ," QoSNET :An integrated QoS network for routing protocols in large scale Wireless sensor networks", Elsevier (2010), Department of Computer Engineering ,Ecole Polytechnique de Montréal

,P.O.Box6079, Centre-Ville Station ,Montreal, Quebec, CanadaH3C3A7

[4] Xiaoxia Huang, Yuguang Fang ," Multiconstrained QoS multipath routing in wireless sensor networks", WirelessNetworks14(2008) 465–478

[5] Gergely Treplan, LongTran –Thanhand Janos Levendovszky," Energy Efficient Reliable Cooperative Multipath Routing in Wireless Sensor Networks", World Academy of Science, Engineering and Technology 68 (2010)

[6] Rajeshwar Singh, Dharmendra K Singh, Lalan Kumar," Performance Evaluation of DSR and DSDV Routing Protocols for Wireless Ad Hoc Networks", Int. J. Advanced Networking and Applications 2011

[7] Md.Abdur RAZZAQUE, Muhammad Mahbub ALAM, Md.MaMUN ,"Multi-Constrained QoS Geographic Routing For Heterogeneous Traffic In Sensor Networks", IEICE TRANS COMMUN VOL.E91–B NO.8 AUGUST (2008)

[8] Asar Ali, Zeeshan Akbar, "Evaluation of AODV and DSR Routing Protocols of Wireless Sensor Networks for Monitoring Applications",Electrical Engineering with emphasis on Telecommunication, Karlskrona October (2009)

[9] Changle Li, Hanxiao Zhang, Binbin Hao and Jiandong Li, "A Survey on Routing Protocols for Large-Scale Wireless Sensor Networks", Sensors (2011), 11, 3498-3526; doi:10.3390/s110403498

[10]ErolGelenbe,PeixiangLiu,BoleslawK.Szymanski,Christopher Morrell," Cognitiveand Self-Selective Routing for Sensor Networks ",Computer Management Science (2010)

[11] Jiang, C.; Yuan, D.; Zhao, Y. Towards clustering algorithms in wireless sensor networks-a survey. In Proceedings of 2009 IEEE Wireless Communication and Networking Conference, Budapest, Hungary, April (2009); pp. 1-6.

[12] Yuan, L.; Wang, X.; Gan, J.; Zhao, Y. A data gathering algorithm based on mobile agent and emergent event-driven in cluster-based WSN. Networks 2010, 5, 1160-1168

[13] Zhou, Y. and Y. Fang, 2008. University of Florida securing wireless sensor networks a survey. IEEE Commun. Surveys Tutorials(2008).

[14] Aslam,N.; Phillips,W. ;Robertson,W. ;Sivakumar, S.Amulti-criterionoptimizationtechnique For energy efficient cluster formation in wireless sensor networks. Inform .Fusion 2009, doi:10.1016/j.inffus.(2009.12.005).

BIOGRAPHIES

ArashGhorbanniaDelavar received the MSc and Ph.D. degrees in computer engineering from Sciences and Research University, Tehran, IRAN, in 2002 and 2007. He obtained the top student award in Ph.D. course. He is currently an assistant professor in the Department of Computer Science, Payam Noor University, Tehran, IRAN. He is also the Director of Virtual University and Multimedia Training Department of Payam Noor University in IRAN. Dr.ArashGhorbanniaDelavar is currently editor of many computer science journals in IRAN. His research interests are in the areas of computer networks, microprocessors, data mining, Information Technology, and E-Learning.

Tayebeh Backtash received the BS, in 2007 and now, She is MS Student in the department of Computer Engineering in Payam Noor University, Tehran, IRAN. Her research interests include computer networks, wireless communication and web programming. She is a high school teacher.

Leila Goodarzi received the BS, in 2006 and now, She is MS Student in the department of Computer Engineering in Payam Noor University, Tehran, IRAN. Her research interests include computer networks, wireless communication and Data mining. She is a high school teacher.

Automated PolyU Palmprint sample Registration and Coarse Classification

Dhananjay D. M.[1], Dr C.V.Guru Rao[2] and Dr I.V.Muralikrishna[3]

[1] Computer Science Department ,JNTU
Hyderabad, Andhra Pradesh, India

[2] Department of Computer Science &Engineering, SR Engineering College,
Warangal, Andhra Pradesh, India

[3] Former director R&D, JNTU
Hyderabad, Andhra Pradesh, India

Abstract

Biometric based authentication for secured access to resources has gained importance, due to their reliable, invariant and discriminating features. Palmprint is one such biometric entity. Prior to classification and identification registering a sample palmprint is an important activity. In this paper we propose a computationally effective method for automated registration of samples from PlolyU palmprint database. In our approach we preprocess the sample and trace the border to find the nearest point from center of sample. Angle between vector representing the nearest point and vector passing through the center is used for automated palm sample registration. The angle of inclination between start and end point of heart line and life line is used for basic classification of palmprint samples in left class and right class.

Keywords: *Average filter, Binarization , Gaussian smoothing, Boundary tracking, Angle between vector, Gradient, Left palm print, Right palmprint.*

1. Introduction

Biometric based personnel authentication has established as robust, reliable methodology. An automated biometric system is based on using invariant physiological or behavioral human characteristics for secured access [3]. Biometric trait such as fingerprint, signature, palmprint, iris, hand, voice or face can be used to authenticate a person's claim. Palmprint is one such biometric trait found to poses stable and unique discriminating features. A sample palmprint has many features such as, principle lines, datum point, ridges, delta point and minutiae features [7]. Due to lack availability of standardized palmprint capturing devices most of the research proposals are using PolyU palmprint database as baseline database, to compare and establish test results. The Biometric Research Centre (UGC/CRC) at The Hong Kong Polytechnic University has developed a real time palmprint capture device, and has used it to construct a large-scale palmprint database. The PolyU Palmprint Database contains 7752 grayscale images [9]. All the captured samples of PolyU database are aligned in a specific direction. Providing a computationally efficient method for palm print sample registration and coarse classification of sample palm print in order to reduce computation burden has motivated our paper. In this paper we propose a method for registration of PolyU palmprint database samples and classification into two basic classes. In all the discussion followed, palmprint sample refers to PolyU palmprint database sample.

This paper is organized in five sections. Section 1 is used for introduction. Key features of palmprint and preprocessing of sample palmprint is discussed in section 2. Method to establish boundary of palmprint along with sample center is presented in section 3. The process of finding out the required angle of rotation for palmprint sample alignment and registration is aimed at in section 4. Section 5 is used for finding the angle of inclination of heart line and life line for basic classification, followed by section 6 containing result and discussion.

2. Key features and pre processing of palmprint

2.1 Key features

Principle lines and datum points are regarded as useful palmprint key features and have been used successfully for verification. In addition, other features associated with a palmprint are palmprint geometrical features, wrinkle features, delta point features, ridges and minutiae features. All palmprints contains three prominent lines known as heart line, head line and life

line. They are regarded as principle lines of the palm print. Heart line starts below little finger and ends below the index finger, Head line starts near the thumb region and ends below the heart line origination point. Lifeline encloses the thumb and adjoining region. Region I is an area enclosed by heart line, region II is an area enclosed by lifeline and region III is an area present below the heart line and enclosed by headline. Many a palmprint also contains line originating near wrist and dividing the head line and marching towards heart line. This line is known as fortune line. Fig.-1 shows a sample of palmprint and key aspects associated with it.

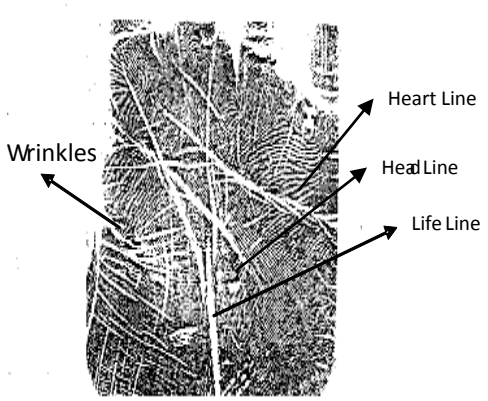


Fig1. Palmprint sample

2.2 Preprocessing of palm print sample

Submitted palmprint sample is submitted for preprocessing. This activity is used to smooth the given sample, obtain binarized sample and also to remove some noise present as additional objects in image. The result of this process is depicted in fig 2 (a,b). Steps involved in this process are as given below

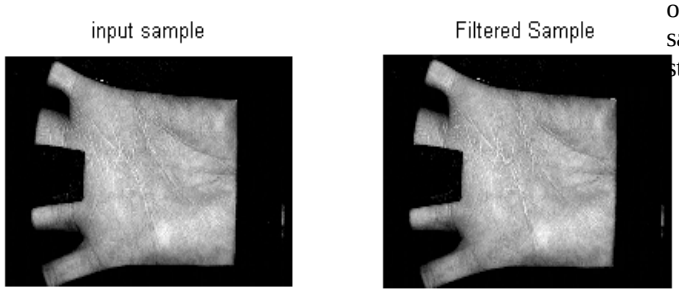


Fig 2a. Sample input and Average filtered sample

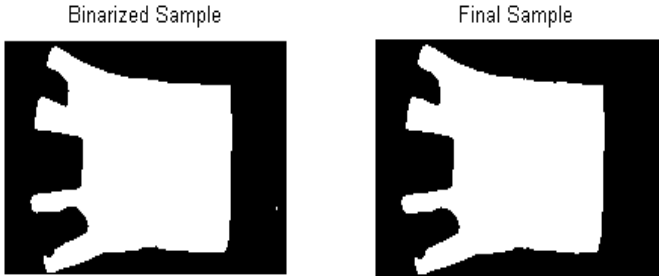


Fig 2b. Binarized image and Final Image

I. Apply 5x5 Gaussian low pass filter mask with standard deviation of value 0.10 to 0.25 on input image. The resulting image is smoothed image [6].

$$IMfilter(i, j) = IM(i, j) \otimes GaussianMask \quad (1)$$

IM (i,j) is input sample it is convolved with Gaussian mask to obtain filters image *IMfilter* (i,j) using (1).

II. Filtered palmprint sample obtained from (1) is submitted for binarization. We select the mean value of the filtered image as threshold and use (2) to convert into binary image.

$$IMbin(i, j) = 1 \; if \; IMfilter(i, j) > \mu$$
$$IMbin(i, j) = 0 \; if \; IMfilter(i, j) < \mu \quad (2)$$
$$\mu \; mean \; of \; \; filtered \; image$$

III. Binarized sample images can contain objects which are of no interest and were retained after filtering. To remove such unwanted objects from image we apply labeling algorithm [8] and calculate area of each labeled objects. Object with largest area is retained as palmprint sample. This process is depicted in fig (2b). Following steps are used for this process.

(i) Obtain labeled image

IMLabel(i,j) = Label(IMbin(i,j)

(ii) Find number of objects with distinct label

Num=(IMlabel==x)

(iii) Calculate area of each object and retain object with maximum area.

(iv) Output image is logical '&'(and) of labeled image

3. **Tracing boundary and establishing nearest point from center**

3.1 Tracing Boundary of palmprint sample.

Output obtained after pre processing the image is used establish the boundary of the image by using suitable boundary tracing algorithm. In this paper we trace the boundary image by first establishing image coordinate with transition from 0 to 1 where 0 represents background and 1 represents object of interest. The neighborhood operation is used to trace the boundary of the image [8]. All co-ordinates representing boundary are collected in a boundary vector using (4). The boundary traced is being represented in fig3.

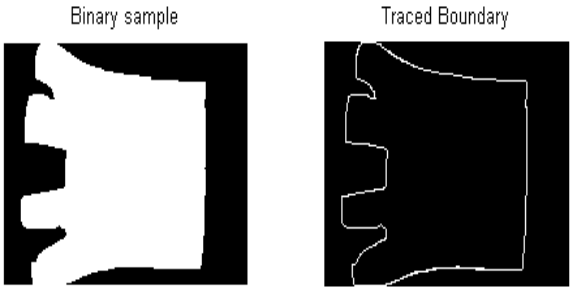


Fig3. Binary sample and traced boundary

$$VectB(i,k) = IMbin(i,j) \;\; if \;\; i,j \in boundary \quad k = 1,2 \tag{4}$$

Further we establish center of the sample by finding of the center of mass of binary image using (5) & (6).

$$X0 = \sum VectB(i,1) / sizeofxcordinate \tag{5}$$

$$Y0 = \sum VectB(i,2) / sizeofxcordinate \tag{6}$$

3.2 Establishing nearest point from center

We use Euclidian distance measure to calculate the distance of points in vector VectB representing sample boundary using (7).

$$dist(i) = EuclidianDist([X0\,Y0], VectB) \tag{7}$$

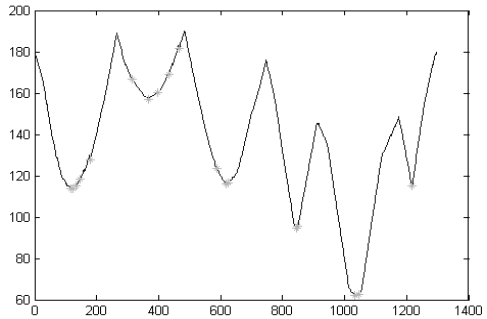


Fig4. Plot representing Euclidian distance from center

A plot of distance from center to bordering elements stored in vector VectB is represented in fig 4. The curve in graph represents curvaceous points of the given sample image. Using curve as an input we locate four points which are at minimum distance from center. Result of this operation is represented in fig 5.

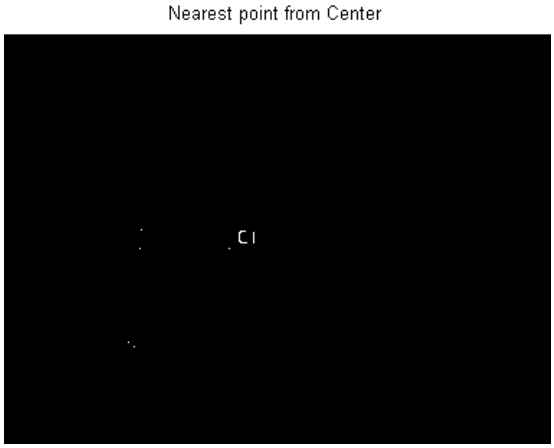


Fig5. Nearest points from center C1

4. **Automated palm print alignment for registration**

Given a vector V1 representing end of a line co-ordinates in Cartesian system and V2 another vector intersecting at point 'C1'. The angle between two vector intersecting at the given point is calculated by arctan of cross product of vector by dot product of vector [5]. We use (8c) to calculate the required angle.

$$V1 \times V2 = \|V1\|\|V2\| Sin\theta \tag{8a}$$

$$V1 \bullet V2 = \|V1\|\|V2\| \cos\theta \tag{8b}$$

$$\theta = \tan^{\varphi}(V1 \times V2)/(V1 \bullet V2) \tag{8c}$$

After establishing the center of palmprint sample, we pass a straight line through C1, this forms vector V1. The four nearest points established through the process discussed in previous section, we find out the nearest point from C1, this will establish the second vector V2. The output of the process is depicted in fig 6.

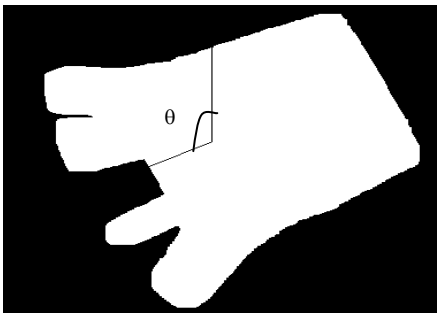


Fig6. Angle θ calculated for sample palm print

The input sample for which we calculated angle θ is rotated for uniform registration. This output is represented in fig7.

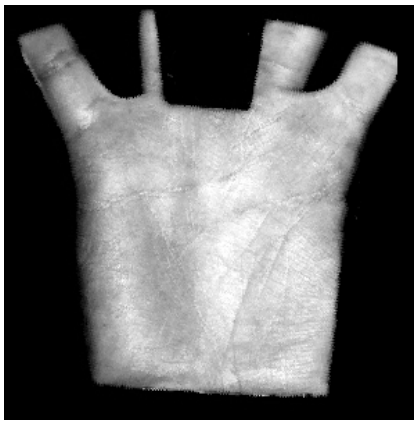


Fig7. Registered Palm print sample

5. Palm print sample course classification

5.1 Heart Line and Life line

Heart line, Head line and life line are considered as principle lines of palm print sample. Head line origination is from below the little finger and end near index finger. If a line is put from start point to end point its inclination can be observed in opposite direction in left a hand and right hand. Life line originates below the thumb region and encircles the thumb region and ends near the wrist. If a line is put from origination point to end points its inclination is in opposite direction for right and left hand. This property is shown in fig 8 using imaginary lines. We use this discriminating property of palm print sample to classify the sample in left & right palmprint sample class

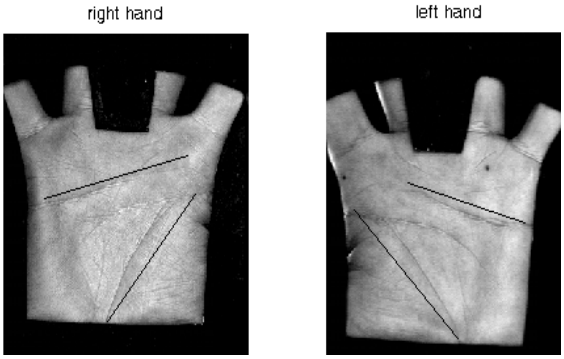


Fig8. Line inclination for left and right palmprint

5.2 Palm print sample classification

If P1 and P2 represent the origination and end point of heart line and Q1, Q2 represent the origination and end point of life line. We calculate the angle of inclination θ1, for line joining P1, P2. Let θ2 represent angle of inclination for points Q1, Q2. A value of +θ1 is present for right hand sample and -θ1 is present for left hand sample. Same holds true for life line using measured angle θ2.

Many of the palmprint sample may contain heart line which of shorter length and horizontal in nature. But all samples will contain life line prominently visible. We use two level checks to decide the class of the sample by using angle θ1 and θ2. If angle θ1 and θ2 are negative then class of sample is right class and if angle θ1and θ2 is positive the class of sample is Left class. The complete process of sample registration with coarse classification is given in algorithm-1 as pseudo code

Algorithm-1

```
Sample=read_sample(palm database);
FilterSample=Apply(Gaussian mask on Sample);
M=mean(Sample);
BinSample=Sample>M;
Label(BinSample);
N=numofobject(BinSample);
Sample=N with Max area;
[x y]=center(sample)
VectB=Boundary(sample);
VectD=EcludianDist(x,y to VectB);
[p1 p2 p3 p4]=min(VectD)
Find Min of p1,p2,p3,p4
V1=[p1, [x y]]
V2=[[x y],0];
Theta=atan(V1*V2)/(V1.V2);
Sample=Rotate(Sample,Theta);
[P1 P2]=select(HeartLine);
[Q1 Q2]=select(Life Line);
Theta1=gradient(P1,P2);
Theta2=gradient(Q1,Q2)
If Theta1 & Theta2 <0
Class=Right
Else
Class=Left
Stop
```

6. Test result and conclusion

6.1 Test Results

Algorithm proposed here is implemented using Matlab7.0. Our primary source of palmprint sample database is PolyU database. Though all samples in the database are aligned in same direction, for testing we have applied image rotation by different angles to consolidate results. Fig9. (a, b, c, d, e, f, g) shows the results for subset of sample input

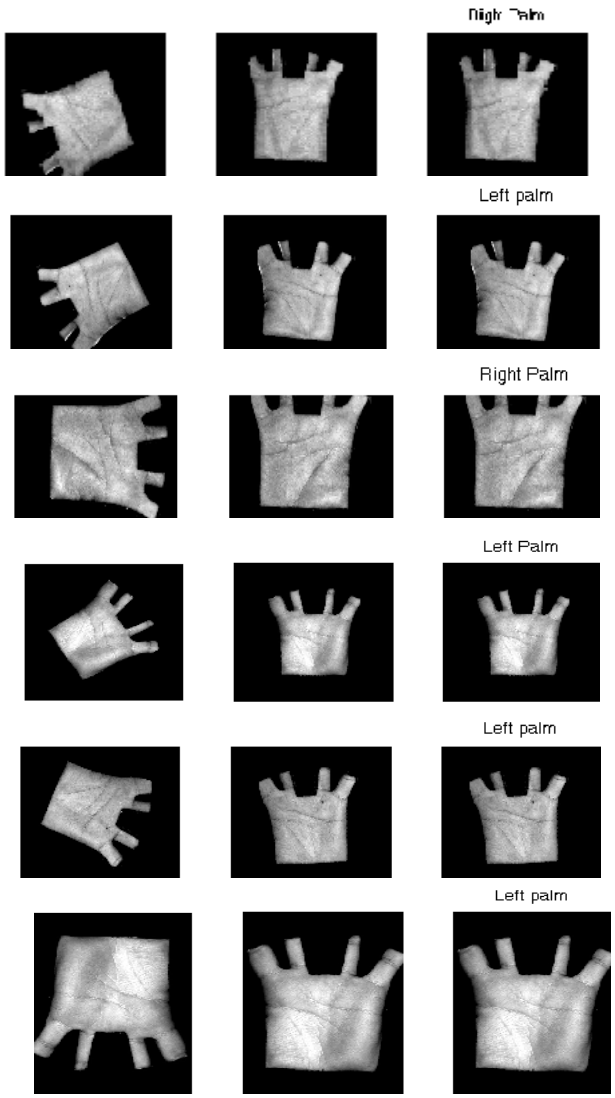


Fig9. (a,b,c,d,e,f) First column Sample input, Second column Aligned sample, Third column classified sample

6.2 Conclusion

Palm print samples archived in PolyU Palm print database are aligned in same direction as capturing device uses peg to restrict the movement. Our proposed algorithm can be used to allow user to obtain palm sample in any direction. The computational burden can be reduced, if primary classification is performed at acquisition level. The method proposed is invariant to transformation, which is useful feature to acquire palmprint sample from other devices.

Acknowledgments

We are thankful to "The Hong Kong Polytechnic University" for (PolyU) Palmprint Database (The Second Version) made available for research. Portions of the work tested on the PolyU Palmprint Database http://www.comp.polyu.edu.hk/~biometrics/

References

[1] H. V. Alexander ,"Classifying Palm Prints", Illinois Charles C. Thomas Publication, 1973

[2] Wei Shu, Gang Rong, Zhaoqi Bian, David Zhang "Automatic Palmprint Verification" International Journal of Image and Graphics, Vol. 1, No. 1 (2001) 135-151

[3] Damien Dessimoz Jonas Richiardi Christophe Champod Dr. Andrzej Drygajlo "Multimodal Biometrics for Identity Documents State-of-the-Art"Research Report PFS 341-08.05 sept2005

[4] Anil K. Jain and Meltem Demirkus " Latent Palmprint Matching" MSU Technical Report, May 2008

[5] Michael Corral Vector Calculus Schoolcraft College 2008 GNU Free Documentationhttp://www.mecmath.net

[6] N. Duta, A.K. Jain, and K.V. Mardia,"Matching of Palmprint," Pattern Recognition Letters, vol. 23, no. 4, pp.477-485, 2001.

[7] D. Zhang, A.W. Kong, J. You and M. Wong, "Online Palmprint Identification," IEEE Trans. Pattern Anal.Mach. Intell., vol. 25, No. 9, pp. 1041-1050, Sept. 2003.

[8] RC Gonzalez and RE Woods,Steven L Eddings "Digital Image Processing using Matlab" Pearson Education 2004

[9] "PolyU Palmprint Palmprint Database, http://www.comp.polyu.edu.hk/~biometrics/"

Dhananjay D M received his BE degree from GUG and has obtained M.Tech from VTU. Presently pursuing PhD under guidance of Dr C.V.Guru Rao and Dr I.V.Muralikrishna. He has presented paper in three conference and one journal. He is member of IETE,CSI India. His area of interest are Biometrics, Neural network, Pattern recognition.

Dr. Guru Rao C V He has distinguished himself as a teacher-researcher-administrator for more than 26 years in service of the student community and the society at large. Presently He is working as a Professor and Head, Department of Computer Science & Engineering at S R Engineering College, Ananthasagar, Warangal. He received a Bachelor's Degree in E&CEfrom Nagarjuna University, Guntur, India in the year 1981. He is double post graduate i.e., M.Tech in Electronic Instrumentation and M.E in Information Science & Engineering from Regional Engineering College, Warangal and Motilal Nehru Regional Engineering College, Allahabad respectively. He was awarded a Ph.D Degree in CS&E IIT, Kharagpur, India in 2004. He had started his career at Kakatiya Institute of Technology & Science, Warangal, as a Lecturer in Electronics & Instrumentation Engineering in 1985. He had served the institute KITS Warangal in various capacities. He was elevated to the post of Principal of the same institute in 2007. He had served as a Member as well as Chairperson, Board of Studies in Computer Science & Engineering and Information Technology, Kakatiya University, Warangal and other institutions for several times. He is a Member on the Academic Advisory Committee of a Deemed University "Monad University, New Delhi" and Advisor of 21st Century Gurukulam, Kakatiya Unviersity, Warangal. He was nominated as a Member on to Industry Institute Interaction (III) Panel by A.P. State Council of Confederation of Indian Industry (CII), Andhra Pradesh. He was also nominated as a member of Engineering Agricultural Medicine Common Entrance Test (EAMCET) Admissions Committee and Post Graduate Engineering Common Entrance Test (PGECET) Committee in 2010 by Andhra Pradesh State Council of Higher Education, Hyderabad. He had published 53 technical research papers in various journals and conferences. A book titled "The Design and Analysis of Algorithms, 2e" by Anany Levitin was adapted by him in tune to the Indian standards and it was published by Pearson Education India. LAP LAMBERT Academic Publishing AG &

Co., Germany had proposed to publish his Ph.D. thesis "Design-For-Test Techniques for SOC Designs" in a book form. Under his guidance 11 students completed their Master's Dissertations including a scholar from Italy. Further, 25 research scholars are working on part-time basis to acquire a Ph.D. at different universities. He had been steering the "International Journal of Mathematics, Computer Science & Information Technology" in bringing out pragmatic research publications as Editor-in-Chief. Further, contributing to the "International Journal of Computational Intelligence Research and Applications" and "International Journal of Library Science" in publishing peer reviewed research articles as Member on the Editorial Board

Dr I.V.Muralikrishna M Tech(IIT-Madras) PhD (IISc-Bangalore) FIE, FIS, FAPASc, FICDM, MIEEE, FIGU. Has been on board of many prestigious institute. He was instrumental in implementation of many research project on weather modification, Environment and cloud seeding . He is receiver of many prestigious awards in recognition of his research work. He has been instrumental in organizing many international and national workshop , seminars, conferences. He was director of R & D at JNTU Hyderabad. Published 83 papers including 44 papers peer reviewed publications and Research Guidance to about 159 Ph D scholars and Graduate students. As on January 2011 Guided / Co-Guided 24 PhDs and 135 M Tech / MCA/ M Sc / MS in Faculty of Spatial Information Technology Faulty of Environmental Science and Technology Faculties of ECE Faculty of Civil Engg & Water Resources Faculty of Computer Science and Technology Faculty of Management Studies.

Color Features Integrated with Line Detection for Object Extraction and Recognition in Traffic Images Retrieval

Hui Hui Wang[1], Dzulkifli Mohamad[2] and N.A.Ismail[2]

[1] Department Of Software Engineering and Multimedia Computing, Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia

[2] Department of Computer Graphics and Multimedia, Faculty of Computer Science and Information Technology, Universiti Teknologi Malaysia, 81300 Skudai, Johor Bahru, Malaysia

Abstract

This paper proposes a simple object extraction and recognition method with efficient searching for identifying and extracting the objects in a complex scene based on the color features. The background of the images is needed to be extracted and recognized in order to get the object of the interested in the images first. This can be achieved by getting the best separation line between building and road, followed by the interested objects (vehicles) on the road. The vehicle objects are represented by using Minimum Bound Rectangle (MBR) and the vehicle object representative points will be the left bottom coordinate of the MBR. The color of the vehicles will be used as the attributes of the objects. Experiments have been conducted to demonstrate that single and multiple known objects in complex scenes can be extracted by using this approach.

Keywords: *Content based Image Retrieval (CBIR), Object Extraction, Object Recognition.*

1. Introduction

Object detection in arbitrary scenes is an important and challenging research topic in computer vision [1] and object searching in a database of color images is a particular problem of color image retrieval similar to appearance-based object recognition [2]. Retrieving known objects from a complex scene is identifying and recognizing the known objects in the scene and determining the region occupied by these objects. In addition to object recognition and scene interpretation, the applications include associative retrieval, querying image databases with visual data, search and replace operations in multimedia retrieval.

Having an accurate object recognition algorithm especially for natural and complex images in the field of image retrieval is contributing to increase the accuracy in image retrieval besides it used as input for semantic features extraction to reduce the semantic gap in image retrieval.

2. Related Works

In the context of object recognition, one of the most widely used image features is the color histogram. It is robust with respect to distortions, including deformation, translation, rotation and scaling of the object. The processing of the color histogram [3] which characterizes the object requires a segmentation step in order to identify the pixels that represent the object. Since the color vectors of the pixels depend on the illumination, the color histograms of two similar images may be different. Therefore, the comparison between color histograms may fail to recognize the same object is illuminated under different illumination conditions [2]. Ref [4] then improved the color histogram approach and proposed an object recognition technique by analyzing their colors when their images are acquired under different illumination conditions. The chromatic co-occurrence matrices are used to characterize the relationship between the color component levels of neighboring pixels. Their matrices are transformed into adapted co-occurrence matrices that are determined so that their intersection is higher when the two images contain the same object is illuminated under different illumination conditions.

However, the above approaches were designed for the object recognition with images that contained one single object placed on an uniform background or multiple objects observed under uncontrolled illuminations. They are having problem when dealing with natural and complex background (with variety of color exist and color that are scatted). Image background may create confusion in recognizing object classes, the background can also provide useful cues to aid recognition, since many objects tend to occur in particular types of scene [5-7]. The use of color histograms is simple and fast, but it works mainly for non-cluttered scenes or for pre-segmented objects. Besides, the

spatial distributions are not take into considerations for capturing the object that exists in the images and consequently it might always caused inaccurate or false objects detection.

There are several techniques proposed to integrate spatial information with color histograms for better and accurate object recognition. G.Pass et al [8] proposed Histogram refinement based on color coherence vectors. The technique considers spatial information and classifies pixels of histogram buckets as coherent if they belong to a small region and incoherent. Huang [9] proposes color correlogram for histogram refinements. Hsu et al. [10] integrated spatial information with color histograms by first selecting a set of representative colors and then analyzing the spatial information of the selected colors using *maximum entropy* quantization with event covering method. Ref M.Stricker et al. [11] partition an image into 5 partially overlapping, fuzzy regions, extract the first three moments of the color distribution for each region, and then organize them into a feature vector of small dimension. Smith and Chang [12] apply back-projection on binary color sets to extract color regions. Vinh Hong et al [13] proposed a color-based object classification. A color histogram is determined by the histogram type (e.g. relative vs. absolute), the color space such as RGB, and the quantization of the color space. The quantization is the process of partitioning the color space into disjoint sub spaces. It uses the nearest neighbour classifier (NN) which is based on a set of classified sample patterns representation of color histogram

Even though the color spatial features has been taken into consideration for better object identification for the above approaches [8-12], however, the object recognition using low level color features is still not fully addressed. In fact, current color features object recognition approaches are used either for Histogram Color Matching in term of similarity then evaluating the similarity between the scene histogram and model histograms or segmenting the image to capture the local features. Hence, it failed to reliably detect and recognize the object. Color Object Classification method [13] also classified object into certain categories based on their color similarity. So object recognition based on color features are needed instead of object matching similarity or object classification.

3. Proposed Solution

In this paper we proposed a strategy for identifying known objects and demarcating the approximate regions occupied by these objects in a complex scene using only color features. The proposed color object recognition can be divided into 2 main stages: Color Features Extraction and Object Identification and Recognition.

3.1 Color Feature Extraction

The Color Features Extraction process is used to extract the features of the images for object recognition purpose. The Color Features Extraction process has five stages and the data flow is shown at Fig.1

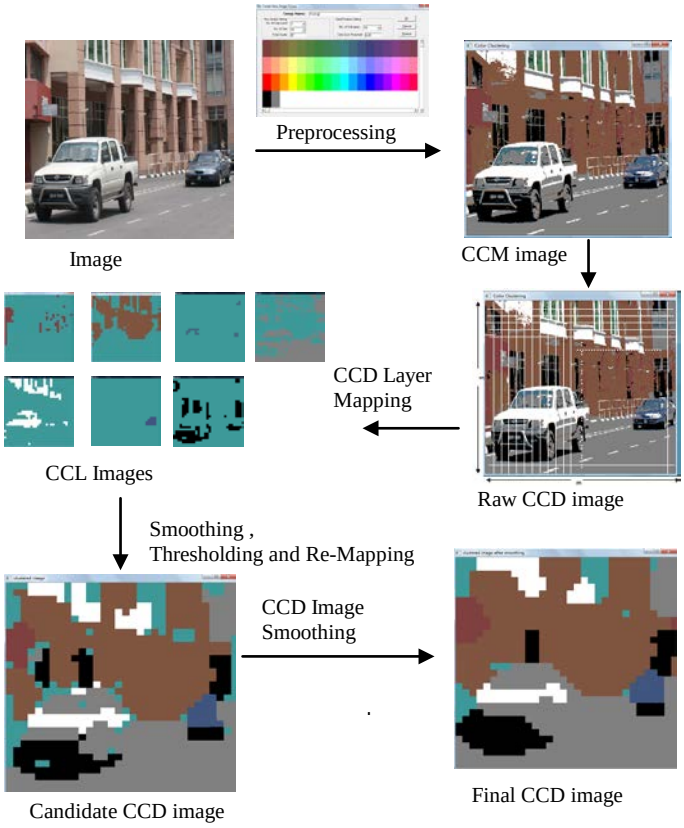


Fig.1: Feature extraction process data flow

3.1.1 Preprocessing

All images need to go through the preprocessing process where it needs to determine the color value of every pixel in the image and compute its distances to all the colors in the predefined color table. The pixel is assigned to the cluster color in the color table that has the smallest distance to the pixel. The output of the preprocessing process produces a Colour Cluster Mapping Image (CCM Image).

3.1.2 Image Sub Division

The CCM Image from the preprocessing process is divided equally into *m x m* sub areas. All the images go through the Image Sub Division process for the purpose of calculating

the population of every colour that exists in each sub area. This process produces a raw Colour Cluster Distribution image (CCD Image).

3.1.3 CCD Layer Mapping

The raw CCD image is going to use for the CCD layer mapping process where the raw CCD image will be extracted and represented into *n* Color Cluster Layer images (CCL Images), where *n* is the number of colours**.** The CCL images then will be used as input for colour CCL smoothing process.

3.1.4 Smoothing, Thresholding and Remapping to CCD Image

All of the CCL images go through the Smoothing process to reduce the noise of the images. The smoothed CCL images will be used for thresholding process. The threshold value is used to identify the population of the color in the image that are needed to be removed as it is consider as a noise of the image. Next, the list of n candidate CCL images will be produced.

The regions that are smaller than an area threshold will be removed from the final segmentation where the results contain only contiguous sets of pixels that have a relatively uniform color distribution and the population of colours are large enough.

All candidate CCL images will be re-mapping and produce another image containing only the dominant clusters, namely Color Cluster Distribution Image (CCD Image).

3.1.5 CCD Image Smoothing

Same as CCL images, the CCD image will go through the CCD image smoothing. Window operation is applied to remove unnecessary noise to come out with a finalized smooth CCD image. The window operation in CCM image is different from CCL image where this CCM is based on few colors in image while CCL is each single color versus to background colour but the concept are the same. The output will be the Final CCD image.

3.2 Object Identification and Recognition

The Final CCD image will be used as input for Object Identification and Recognition process. Object Identification and recognition is needed to extract the object of interest in the images. The Object identification and recognition process has 9 stages and the data flow is shown at Fig.2

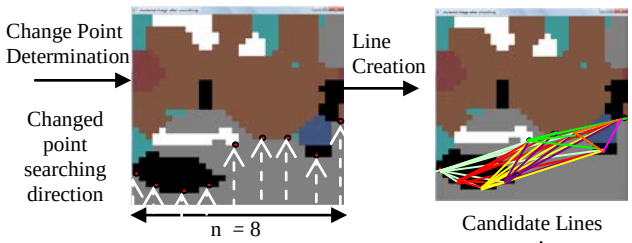


Fig 2. Object Identification and Recognition

3.2.1 Changed Point Determination

The final CCD image will be used for road and building extraction in the traffic images. Firstly, it needs to go into changes point determination process to determine the color changes point. The image is divided into m size at axis-x and the starting point is from the bottom left of the images. The changed point will be the first changes of color area based on the axis-x. This process is continued until get *n* of points, n is the number of changing points that predefined.

3.2.2 Line Creation

The road slope is used as the reference slope for getting the object of interest (vehicles). In the line creation process, all possible lines will be created based on the n points from the changed point determination process. List of lines will be created as output form the line creation process as shown in Fig.2

3.2.3 Line Verification

Line verification is needed to get the candidate lines by removing all the negative slope lines. All candidate lines will go into slope of line calculation process to get the value of the slope for each possible pair of line.

Since this research domain images are traffic images with perspective view. So, assumption has been made as below,

1. There is no negative road slope
2. The best slope value (reference slope) is approximate 0.20 (This value is obtained based on human perception judgment from collection of database images)

Given changed points of (x1, y1) and (x2,y2) of the angle between 2 lines, the slope of line, m can be calculated using formula below.

$$\text{Road Slope, m} = \frac{y2 - y1}{x2 - x1}$$

3.2.4 Best Line Drawing

All candidate lines with their slope values will be compared with the slope reference value. The line with slope value that is nearest to the reference slope value will be chosen as reference slope of the image.

So, the 2 coordinates of the reference slope will be used as 2 reference points to get the best line of the road that act as a z-axis due to the image view is slanted.

3.2.5 Color Cluster Group Identification

All Color cluster groups are identified based on all the color representatives that are available in the images. For each color cluster, two coordinates from axis-x and axis-y which are the nearest and farest from axis-x and axis y respectively will be chosen as the representative coordinate to form a color minimum bound rectangle that is used to represent the color cluster group.

3.2.6 Horizontal and Vertical Scanning

All the color cluster groups will be divided into an Individual Color Clusters (represented by red dotted box) by the horizontal and vertical scanning process.

3.2.7 Object Cluster Identification

The Individual color clusters will be used as an input for the object cluster identification process to get the candidate object clusters. The Candidate object clusters are classified into two groups which are region/object clusters and Reference object clusters. Since the object of interest is a vehicle. So assumption need to make that all vehicles will have black tyre and most of the time, there is a shadow below the vehicle. So, the black colour cluster will be used as reference objects to search for candidate object.

3.2.8 Object Cluster Verification

Object Cluster Verification is the process to verify the candidate objects. The object cluster and reference object cluster will be combined if they satisfied the predefined object distance and the road distance value. The combination of reference object cluster with object cluster will formed the list of objects clusters.

3.2.9 Object Identification

All object clusters (vehicle objects) will be represented by a MBR and left bottom of MBR will be used as reference coordinate of the vehicle objects and the color of the vehicle will be used as attributes for the car.

4. Experiments

The Objective of the experiment is to evaluate the accuracy and effectively of the proposed color object recognition method in recognizing the objects of interest (vehicles) in the domain of traffic images.

4.1 Experiment Setting

There are 3 parameters setting in the experiments which are the number of colour cluster that used as the colour representation of the image, the color cluster object distance value used for determining the value distance between color clusters are considered to be merged and also the road distance value that used for determining the distance value from the road slope to the color cluster are considered to identify as object as interest. In these experiments, there are 57 color cluster used as color representation of the image, color cluster object distance of 3 and road distance value of 2.

4.2 Experiments

To measure the accuracy of the object detection and recognition, the detected vehicles by the proposed color feature object recognition compared with vehicle outlines selected by a user on the same images. A successful detection is considered as detected and correct if the detected vehicles by the proposed algorithms are same as the vehicles selected and defined by users. If the prototype missed the detected cars that user defined, it is consider as missed car. For false detection, the prototype detected the car which is not defined by the users. Two experiments were carried out for single and multiple vehicles detection and recognition in complex and natural images.

4.3 Results and Analysis

The results of the experiments are summarized and shown in Table 1.

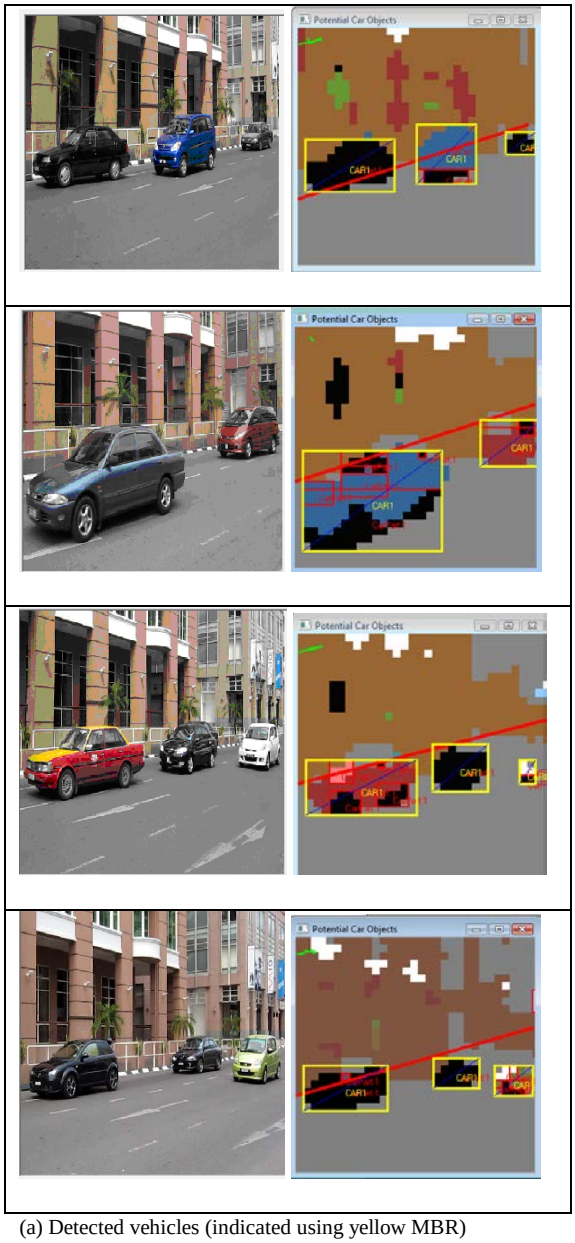
Experiments	Total vehicles	Detected vehicles	Missed vehicles
1 (single vehicles only)	100	97	2
2 (2 vehicles and above)	120	108	8

False Detection	Detection rate
1	97%
4	90%

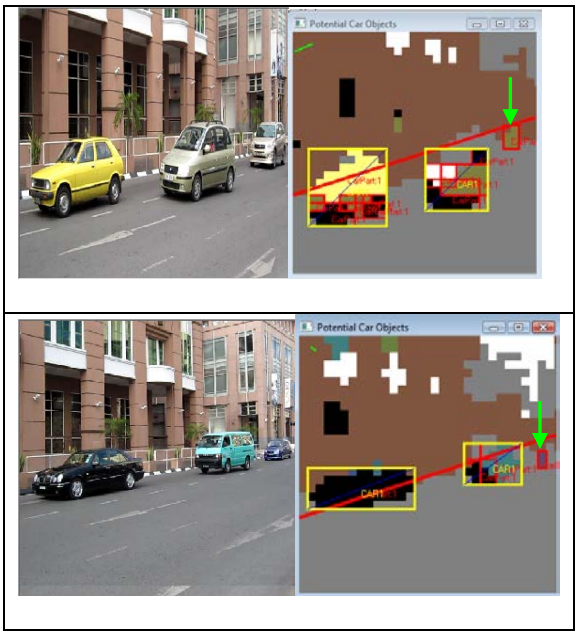
Table 1. The experiment results of the proposed color features object recognition.

The proposed method has proven to be successful in detecting the vehicles. The results accuracy is 97% and 90% respectively for the detection of single and multiple vehicles in the images. So the proposed vehicle detection method is accurate and robust under complex and natural background and it also supports multiple vehicles detection and recognition. This technique has been used as input for Semantic object spatial relationships extraction and representation in our paper [14] in the area of Semantic based image retrieval.

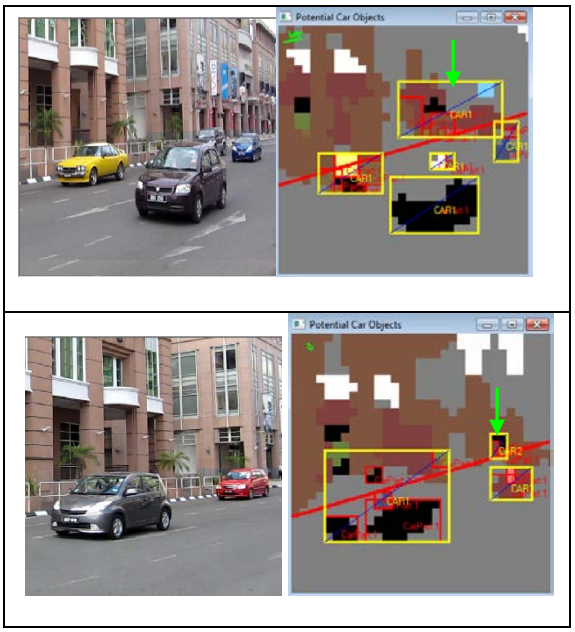
Some of the object detection and recognition results from proposed method is shown in Fig. 3



(a) Detected vehicles (indicated using yellow MBR)



(b) Missed Car(s) (pointed by Green Arrow)



(c) False Detection (pointed by Green Arrow)

Fig. 3 Some results of the experiments

There are some missed vehicles detection (Fig.3(b)) due to the noise created from the smoothing process and also unable to get the black cluster as reference object. There are some false detection (Fig.3(c)) due to the variety color representative of the cars besides some noise created.

4. Conclusions and Future works

In conclusion, a simple object extraction and recognition method with efficient searching for identifying and extracting the objects in a complex scene based on the color features has been proposed and developed. Experiments have been carried out and it is proved that the proposed method works well in detecting both single and multiple objects in natural and complex background for traffic image retrieval. This method is designed especially for use in the automatic semantic object spatial relationship extraction and representation that designed in Ref [14]. This object detection method can be further improved by determining the type of the vehicles.

References

[1]Chia-FengJuang,Wen-KaiSun,Guo-CyuanChen.“Object detection by color histogram based fuzzy classifier with support vector learning”, Neurocomputing 2009, pp.2464–2476
[2] Damien Muselet, Ludovic Macaire. ‘Combining color and spatial information for object recognition across illumination changes’. Pattern Recognition Letters 28 (2007) pp.1176–1185
[3] M. Swain and D. Ballard. “Color indexing”, International Journal of Computer Vision, 7(1), 1991, pp. 11–32.
[4] Damien Muselet, Ludovic Macaire. “Combining color and spatial information for object recognition across illumination changes”. Pattern Recognition Letters 28 (2007) 1176–1185
[5] J. Lim, P. Arbeláez, C. Gu, and J. Malik. “Context by Region Ancestry”. In Proceedings of the IEEE International Conference on Computer Vision, 2009.
[6] J. Uijlings, A. Smeulders, and R. Scha . “What is the spatial extent of an object?” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[7] Yu Su, Moray Allan, Frederic Jurie, Greyc. “Improving object classification using semantic attributes”. BMVC 2010
[8] G. Pass and R. Zabih. “Histogram Refinement for Content Based Image Retrieval”, 3rd IEEE Workshop on Applications of Computer Vision, WACV , 1996 , pp. 96- 102.
[9] J. Huang “Color-Spatial Image Indexing and Applications”. PhD thesis, Cornell Univ., 1998
[10]W. Hsu, T.S. Chua, and H. K. Pung. “An integrated color-spatial approach to content-based image retrieval”. Proceedings of the third ACM international conference on Multimedia 1995
[11] M. Stricker and A. Dimai. “Color indexing with weak spatial constraints”. SPIE proceedings, 2670:29 – 40, February 1996
[12] J. Smith and S.-F. Chang. “Tools and Techniques for Color Image Retrival”. SPIE proceedings, pages 1630 – 1639, 1996.
[13] Vinh Hong Paulus, D. “Parameter Study and Optimization of a Color-Based Object Classification System” IEEE 2009
[14] Hui Hui Wang, Dzulkifli Mohamad, N.A.Ismail. “Semantic Gap in CBIR: Automatic Object Spatial relationships Semantic Extraction and Representation”. International Journal Of Image Processing (IJIP), 2010 Volume (4) : Issue (3)

H.H.Wang received the Bachelor of Information Technology and the Master of science from Universiti Malaysia Sarawak (UNIMAS) in 2001 and 2003 respectively. Her master was on Content Based Image Retrieval. Her intelligent image finder project won the gold metal, Swiss government special award for information solution from 34th Geneva international invention exposition, 2006. She served as a lecturer at UNIMAS and she is now a Ph.D candidate in Department of Computer Graphics, Faculty of Computer Science and Information Technology, Universiti Teknologi Malaysia (UTM). Her main research interests are in image retrieval, computer vision and Pattern recognition.

Dzulkifli Mohamad completed his degree in Computer Science at Universiti Kebangsaan Malaysia and Advanced Diploma in Computer Science at Glasgow University, Scotland. He continued his master's degree and PhD in Computer Science both at Universiti Teknologi Malaysia. .He is currently a professor in the Faculty of computer science and Information Technology at Universiti Teknologi Malaysia. His research interests include areas such as Artificial Intelligence and Identification System

N.A. Ismail received his B.Sc. from Universiti Teknologi Malaysia, UTM, Master of Information Technology (MIT) from National University of Malaysia, and Ph.D. in the field of Human Computer Interaction (HCI) from Loughborough University. He has been a lecturer at Computer Graphics and Multimedia Department, Universiti Teknologi Malaysia for about thirteen years and currently, he is Research Coordinator of the Department. He has made various contributions to the field of Human Computer Interaction (HCI) including research, practice, and education.

Comprehensive Analysis of Web Log Files for Mining

Vikas Verma [1], A. K. Verma [2], S. S. Bhatia [3]

[1] M. M. Institute of Computer Tech. & Business Management , M. M. University, Mullana, Ambala, Haryana-133203, India.

[2] Dept. of Computer Sc. and Engg., TIET, Thapar University, Patiala, Punjab-147004, India.

[3] School of Mathematics and Computer Applications, Thapar University, Patiala, Punjab-147004, India.

Abstract

World Wide Web is a global village and rich source of information. Day by day number of web sites and its users are increasing rapidly. Information extracted from WWW may sometimes do not turn up to desired expectations of the user. A refined approach, referred as Web Mining, which is an area of Data Mining dealing with the extraction of interesting knowledge from the World Wide Web, can provide better result. While surfing the web sites, users' interactions with web sites are recorded in web log file. These Web Logs are abundant source of information. Such logs when mined properly can provide useful information for decision making. Mining of these Web Logs is referred to as Web Log Mining. This paper analyses web log data of NASA of the month of August 1995 of 15.8MB and depicts certain behavioral aspects of users using web log mining.

Keywords: *Web Mining, Data Mining, Web Log Mining.*

1. Introduction

The expansion of the World Wide Web has resulted in a large amount of data that is now in general freely available for user access. The different types of data have to be managed and organized such that they can be accessed by different users efficiently. Therefore, the application of data mining techniques on the Web is now the focus of an increasing number of researchers. Several data mining methods are used to discover the hidden information in the Web. However, Web mining does not only mean applying data mining techniques to the data stored in the Web. The algorithms have to be modified such that they better suit the demands of the Web. In accordance with Kosala, Blockeel and Neven [1], the term 'Web Mining' is defined as the whole of data mining and related techniques that are used to automatically discover and extract information from web documents and services. Web mining research, is an integrate research from several research communities such as: Database (DB), Information retrieval (IR), The sub-areas of machine learning (ML) and Natural language processing (NLP).

An important constituent category of Web Mining is Web Log mining also known as Web Usage mining, is the process of extracting interesting patterns from web access logs [6]. The different techniques are represented through Figure 1. However, not much concentration is done on techniques, since the focus of this paper is exclusively on web logs. Web usage data can include a variety of data from different sources. These sources can include web server access logs, proxy server logs, browser logs or any other data that is generated by users interacting with a website. The issues are outlined by Linoff and Berry [3].

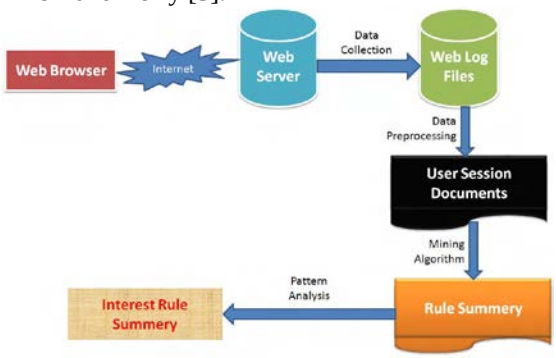


Fig. 1: Web Mining Techniques

There are several general challenges associated with obtaining due results from the data. Firstly, extraneous information is mixed with useful one. Secondly, multiple server requests may be generated by a single user action. Thirdly, multiple user actions may generate the same server request. Fourthly, local activities (for example browser navigation using 'back', and 'forward' buttons) are not recorded.

The paper is organized as follows: In section 2, we emphasize on web logs; in section 3, latest developments in the field web usage mining are presented; in section 4, visualization is depicted through a tool; section 5, focus on the future prospects and conclusion.

II. Motivation

Web Log mining is the process of identifying browsing patterns by analyzing the user's navigational behavior. A Web log file [4] records activity information when a Web user submits a request to a Web Server. The main source of raw data is the web access log. Web server logs are plain text (ASCII) files, independent of server platform. There are some differences between server

software, but traditionally there are four types of server logs [5, 14]:

1. Transfer (access) log
2. Error log
3. Referrer log
4. Agent log

The first two types of log files are "standard / common". The referrer and agent logs may or may not be "turned on" at the server or may be added to the transfer log file to create an "extended" log file format. Each HTTP protocol transaction, whether completed or not, is recorded in the logs, and some transactions are recorded in more that one log. For example, most (but not all) HTTP errors are recorded in the transfer log and the error log.

A transfer access log typically is a long line of ASCII text, separated by tabs and spaces. A sample log is considered below.

1Cust216.tnt1.santa-monica.ca.da.uu.net - - [17/Sept/2011:12:13:03 -0700] GET /gen/meeting/ssi/next/HTTP/1.0 200 9887 http://www.slac.stanford.edu/ Mozilla/3.01-C-MACOS8 (Macintosh; I; PPC) GET /gen/meeting/ssi/next/ - HTTP/1.0

An analysis of each section is done as below

- *1Cust216.tnt1.santamonica.ca.da.uu.net*

This is the address of the computer making the HTTP request. The server records the IP and then, if configured, will lookup the Domain Name Server (DNS).

- *RFC931 (or identification)*

Rarely used, the field was designed to identify the requestor. If this information is not recorded, a hyphen (-) holds the column in the log.

- *Authuser -*

List the authenticated user, if required for access. This authentication is sent via clear text, so it is not really intended for security. This field is usually filled by a hyphen (-).

- *Time Stamp*

[17/Sept/2011:12:13:03 -0700]

The date, time, and offset from Greenwich Mean Time are recorded for each hit. The date and time format is: DD/Mon/YYYY HH:MM:SS. The example shows that the transaction was recorded at 12:13 pm on Sept 17, 2011 at a location 7 hours behind GMT.

- *Request*

GET /gen/meeting/ssi/next/ HTTP/1.0

One of three types of HTTP requests is recorded in the log. GET is the standard request for a document or program. POST tells the server that data is following. HEAD is used by link checking programs, not browsers, and downloads just the information in the HEAD tag information. The specific level of HTTP protocol is also recorded.

- *Status Code 200*

There are four classes of codes

1. Success (200 series)
2. Redirect (300 series)
3. Failure (400 series)
4. Server Error (500 series)

A status code of 200 means the transaction was successful. Common 300-series codes occurs when the server checks if the version of the file or graphic already in cache is still the current version and directs the browser to use the cached version. The most common failure codes are 401 (failed authentication), 403 (forbidden request to a restricted subdirectory), and the dreaded 404 (file not found) messages.

- *Transfer Volume 9887*

For GET HTTP transactions, the last field is the number of bytes transferred. For other commands this field will be a hyphen (-) or a zero (0). The transfer volume statistic marks the end of the common log file. The remaining fields make up the referrer and agent logs, added to the common log format to create the "extended" log file format. An analysis of each section is done as below

- *Referrer URL*

http://www.slac.stanford.edu/

The referrer URL indicates the page where the visitor was located when making the next request. The actual request is shown in the last field of the entry

GET /gen/meeting/ssi/next/ - HTTP/1.0 and is duplicated from the HTTP Request.

- *User Agent*

Mozilla/3.01-C-MACOS8 (Macintosh; I; PPC)

The user agent is information about the browser, version, and operating system of the reader.

The description can be generalized through the following table

Table 1: Web log file attributes and their description

Attributes	**Description**
Client IP	Client Machine IP Address
Client Name	Client Name if required by server, otherwise, hyphen
Date	Date when user made access
Time	Time of transaction
Server Site Name	Internet service name as appeared on client machine
Server Computer Name	Server Name
Server IP	Server IP provided by Internet Service Provider
Server Port	Server port configured for data transmission
Client Server Method	Client Method or modes of request can be GET, POST of HEAD
Client Serves URI Stem	Targeted default web page of web site
Client Server URI Query	Client query which starts after "?"
Server Client Status	Status Code returned by the server like 200, 404
Server Client win32Status	Windows status code
Server Client Bytes	Number of bytes sent by server to client
Client Server bytes	Number of bytes received by Client
Time Taken	How much spend by client to perform any action
Client Server Version	Protocol version like HTTP
Client Server Host	Host header name
User Agent	Browser type that client used
Cookies	Contents of cookies
Referrer	Link from where client jump to this site

III. Work Done

A lot of research projects deal with the Web Log mining. Data Preprocessing, Pattern discovery, and pattern analysis are considered as important phases of Web Log mining process. Most of the efforts focus on extracting useful patterns and rules using data mining techniques in order to understand the users' navigational behavior. Much of the work in this field focuses on user identification, session identification etc. Specifically for web log files [6, 7] has explored certain issues regarding web server log files. Since in Web Log mining several techniques can be used [13], one such technique is Association mining using Web Logs [8, 9 10]. Sequence mining, which is another technique, can be used for discover the web pages which are accessed immediately after another. It is used in [11], using a tree for storing patterns efficiently. [13] Discussed the structure of web log file in detail and performed two preprocessing techniques data cleaning and user identification. [15] Derived the user profiles from the analysis of web log file and Meta data of page contents. [16] Identified that web usage profiles play an important role in web personalization. Profiles were extracted from clusters and clusters were extracted from web usage data after preprocessing the web log file. The navigation pattern can be examined with the data of the server log file by the web analyzer [17].

IV. Visualization

Using weblogexpert [12], software for Web Log mining, NASA server log file is analyzed. Typical behavior based on statistical analysis of the log file is observed and thereafter result is visualized as shown in Figure 2, 3 and 4 depicting Daily Visitors, Daily Error types and Activity-day wise for the month of August 1995. It should be observed that these results can be utilized further for certain web specific applications also.

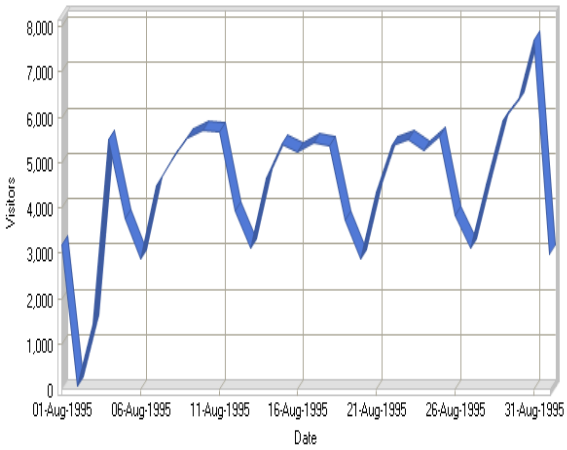


Fig. 2: Daily Visitors schedule

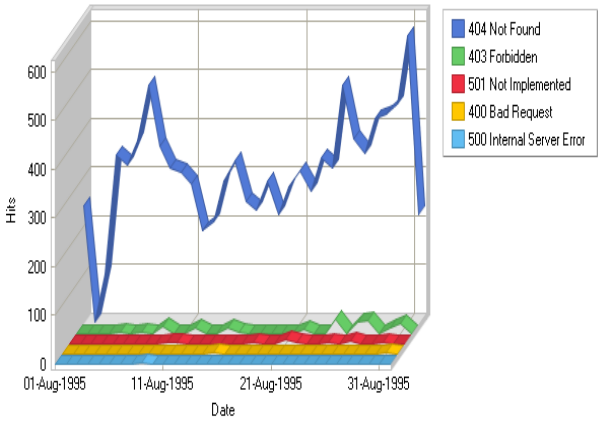


Fig. 3: Daily Error Types

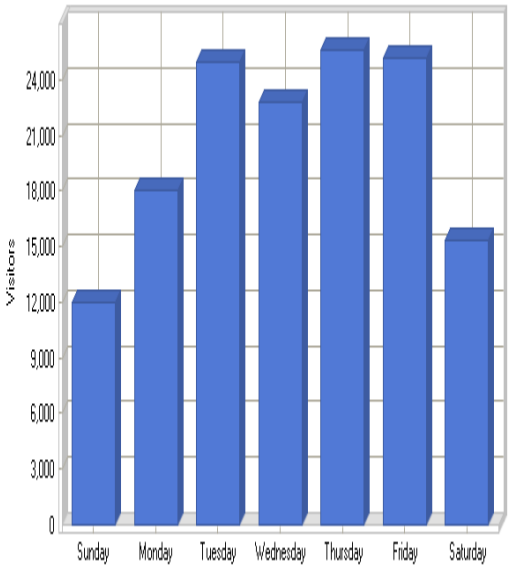


Fig. 4: Activity by Day of Week

V. Conclusions and Scope

The requirement for predicting user needs in order to improve the usability and user retention of a Web site can be addressed by Processing Web Log file efficiently. Future scope of Web Log mining is in Web Personalization and to improve the overall performance of future accesses. In today's era of advancements it can also be used in e-commerce, digital libraries etc, using techniques of data mining at group level instead at individual level for high accuracy.

References

[1] R. Kosala, and H. Blockeel, "Web Mining Research: A Survey" , SIGKDD Explorations, 2(1):1-15, 2000.
[2] O.R. Zaıane, "Building Virtual Web Views", Data and Knowledge Engineering , 39:143–163, 2001.
[3] G.S. Linoff, and M.J.A. Berry, Mining the Web, John Wiley and Sons, first edition, 2001.
[4] Magdalini Eirinaki, and Michalis Vazirgiannis, "Web Mining for Web Personalization", PKDD, 2005.

[5] Internet: Web Log files overview: http://www.si.umich.edu/Classes/540/Readings/ServerLogFileAnalysis.htm

[6] Zhang Huiying, and Laing Wei, "An Intelligent Algorithm of Data Pre-processing in Web Usage Mining ", Proceedings of the 5th world Congress on Intelligent Control and Automation, June15-19, 2004 , Hangzhou, P.R.China.

[7] Doru Tanasa et.al., "Advanced data preprocessing for inter sites Web Usage mining", IEEEE computer society, 2004.

[8] M. Eirinaki, and M. Vazirgiannis, "Web mining for web personalization", ACM Trans. Inter. Tech., vol. 3, no. 1, pp. 1-27, 2003.

[9] J. Punin, M. Krishnamoorthy, and M. Zaki, "Web usage mining: Languages and algorithms", in Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag, 2001.

[10] P. Batista, M. ario, and J. Silva, "Mining web access logs of an on-line newspaper", NetLab, Lund University Libraries, Sweden April 2002.

[11] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu, "Mining access patterns efficiently from web logs", in PADKK '00: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications, London, UK: Springer-Verlag, 2000, pp. 396-407 .

[12] Internet: Typical softwares : http://www.kdnuggets.com/software.html

[13] Suneetha, K. R. and D. R. Krishnamoorthi , "Identifying User Behavior by Analyzing Web Server Access Log File", in IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.

[14] M. H. A. Wahab, M. N. H. Mohd, et al. , " Data Preprocessing on Web Server Logs for Generalized Association Rules Mining Algorithm", World Academy of Science, Engineering and Technology, 2008.

[15] G. Stermsek, M. Strembeck, et al. , " A User Profile Derivation Approach based on Log-File Analysis", IKE 2007, pp. 258-264.

[16] G. Castellano, F. Mesto, et al. , " Web User Profiling Using Fuzzy Clustering", WILF 2007, pp. 94–101.

[17] J Vellingiri, and S.Chenthur Pandian, "A Survey on Web Usage Mining", Global Journal of Computer Science and Technology, Volume 11, Issue 4, Version 1.0, March 2011.

Vikas Verma is currently working as a Lecturer in M. M. Institute of Computer Technology & Business Management, M. M. University, Mullana, Haryana. He received his MCA from Punjabi University, Patiala, Punjab, India, in 2003. He is pursuing Ph.D. from School of Mathematics and Computer Applications, Thapar University, Patiala, Punjab, India. He is in teaching since 2003. He has published 08 research papers in International/National Journals and Conferences. His research interests are Web Mining, Knowledge Discovery, Information Systems, Data base management systems.

Dr. A.K. Verma is currently working as an Associate Professor in the department of Computer Science and Engineering at Thapar Institute of Engineering & Technology (Deemed University), Patiala. He received his B.S., M.S. and Ph.D. in 1991, 2001 and 2008 respectively, majoring in Computer science and engineering. He has worked as Lecturer at M.M.M. Engg. College, Gorakhpur from 1991 to 1996. He joined Thapar Institute of Engineering & Technology in 1996 as a Systems Analyst in the Computer Centre and is presently associated with the same Institute. He has been a visiting faculty to many institutions. He has published over 35 papers in referred journals and conferences (India and Abroad). He is a MISCI(Turkey), LMCSI (Mumbai), GMAIMA (New Delhi). He is a certified software quality auditor by MoCIT, Govt. of India. His research interests include wireless networks, routing algorithms and securing ad hoc networks.

Dr. S.S. Bhatia, is currently working as Professor and Head, School of Mathematics and Computer Applications, Thapar University, Patiala, Punjab, India. He received his M.Sc, M.Phil and Ph.D in 1985, 1986 and 1994 respectively, majoring Mathematics. He is associated with Thapar University for the last 24 years. He has vast experience of teaching at UG and PG level to Science and Engineering students at Thapar University. He has published over 55 research papers in Journals, International and National Conferences in the areas of Functional Analysis, Reliability Analysis and Image processing. He has guided 2 Ph.D. and 7 M.Phil scholars. He has accomplished 2 UGC Major Research Projects. He's a Life member of Punjab Academy of Sciences, Indian Society of Industrial and Applied Mathematics and Indian Society of Technical Education.

Efficient Web Usage Mining with Clustering

K.Poongothai[1] M.Parimala[2] and Dr. S.Sathiyabama[3]

[1] Asst Prof, Department Of Information Technology, Selvam College of Technology, Namakkal, Tamilnadu, India.

[2]Lecturer, Department Of MCA, M.Kumarasamy College of Engg, Karur, Tamilnadu, India.

[3]Assistant Professor of Computer Science, Thiruvalluvar Govt Arts and science college, Rasipuram, Tamilnadu, India.

Abstract

Web usage mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the Web. Web usage mining has become very critical for effective Web site management, creating adaptive Web sites, business and support services, personalization, network traffic flow analysis etc., Web site under study is part of a nonprofit organization that does not sell any products. It was crucial to understand who the users were, what they looked at, and how their interests changed with time. To achieve this, one of the promising approaches is web usage mining, which mines web logs for user models and recommendations. Web usage mining algorithms have been widely utilized for modeling user web navigation behavior. In this study we advance a model for mining of user's navigation pattern.

The proposal of our work proceeds in the direction of building a robust web usage knowledge discovery system, which extracts the web user profiles at the web server, application server and core application level. The proposal optimizes the usage mining framework with fuzzy C means clustering algorithm (to discover web data clusters) and compare with Expected Maximization cluster system to analyze the Web site visitor trends. The evolutionary clustering algorithm is proposed to optimally segregate similar user interests. The clustered data is then used to analyze the trends using inference system. By linking the Web logs with cookies and forms, it is further possible to analyze the visitor behavior and profiles which could help an e-commerce site to address several business questions. Experimentation conducted with CFuzzy means and Expected Maximization clusters in Syskill Webert data set from UCI, shows that EM shows 5% to 8% better performance than CFuzzy means in terms of cluster number.

1. Introduction

Web mining is the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, us-age logs of web sites, etc. Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to web content has been the most widely researched. Issues addressed in text mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages.

Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision, the application of these techniques to web content mining has been limited. The structure of a typical web graph consists of web pages as nodes, and hyper-links as edges connecting related pages. Web structure mining is the process of discovering structure information from the web. This can be further divided into two kinds based on the kind of structure information used.

A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. A hyperlink that connects to a different part of the same page is called an intra-document hyperlink, and a hyperlink that connects

two different pages is called an inter-document hyperlink. In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications. Usage data captures the identity or origin of web users along with their browsing behavior at a web site. Web usage mining itself can be classified further depending on the kind of usage data considered i.e web server data, application server data and application level data.

In Web Server Data, user logs are collected by the web server and typically include IP address, page reference and access time. In Application Server Data, commercial application servers such as Web logic, Story Server, have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs. In Application Level Data, new kinds of events can be defined in an application, and logging can be turned on for them, generating histories of these events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the above the categories.

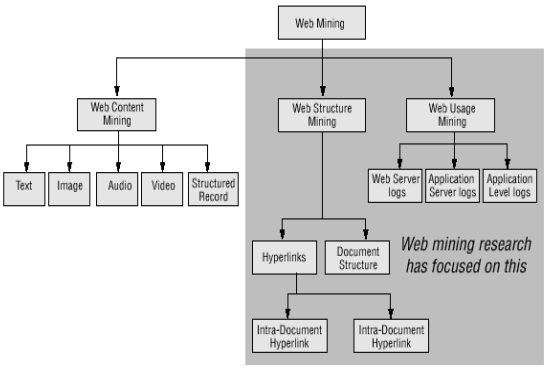


Fig 1: Web Mining Taxonomy

2) Literature Review

The WWW continues to grow at an amazing rate as an information gateway and as a medium for conducting business. Web mining is the extraction of interesting and useful knowledge and implicit information from artifacts or activity related to the WWW [4], [6]. Based on several research studies we can broadly classify Web mining into three domains content , structure and usage mining [8], [9]. This work is concerned with Web usage mining. Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the Web access logs can help understand the user behavior and the web structure. From the business and applications point of view, knowledge obtained from the Web usage patterns could be directly applied to efficiently manage activities related to e business, e-services, e-education and so on [10]. Accurate Web usage information could help to attract new customers, retain current customers, improve cross marketing/sales, effectiveness of promotional campaigns, track leaving customers and find the most effective logical structure for their Web space [3]. User profiles could be built by combining users' navigation paths with other data features, such as page viewing time, hyper- link structure, and page content [12]. What makes the discovered knowledge interesting had been addressed by several works [11] and [12]. Results previously known are very often considered as not interesting. So the key concept to make the discovered knowledge interesting will be its novelty or unexpected appearance.

Whenever a visitor accesses the server, it leaves the IP, authenticated user ID, time/date, request mode, status, bytes, referrer, agent and so on. The available data fields are specified by the HTTP protocol. There are several commercial software that could provide Web usage statistics[1]. These stats could be useful for Web administrators to get a sense of the actual load on the server. However, the statistical data available from the normal Web log data files or even the information provided by Web trackers could only provide the information explicitly because of the nature and limitations of the methodology itself. Generally, one could say that the analysis relies on three general sets of information given a current focus of attention past usage patterns, degree of shared content and inter-memory associative link structures. After browsing through some of the features of the best trackers available it is easy to conclude that rather than generating statistical data and texts they really do not help to find much meaningful information.

For small web servers, the usage statistics provided by conventional Web site trackers may be adequate to analyze the usage pattern and trends. However as the size and complexity of the data increases, the statistics provided by existing Web log file analysis tools may prove inadequate and more intelligent knowledge mining techniques will be necessary[2], [3]. In the case of Web mining, data could be collected at the server level, client level, proxy level or some consolidated data. These

data could differ in terms of content and the way it is collected etc. The usage data collected at different sources represent the navigation patterns of different segments of the overall Web traffic, ranging from single user, single site browsing behavior to multi-user, multi-site access patterns. Web server log does not accurately contain sufficient information for inferring the behavior at the client side as they relate to the pages served by the Web server.

To demonstrate the efficiency of the proposed frameworks, Web access log data at the Monash University's Web site were used for experimentations. The University's central web server receives over 7 million hits in a week and therefore it is a real challenge to find and extract hidden usage pattern information. To illustrate the University's Web usage patterns, average daily and hourly access patterns for 5 weeks are shown. The average daily and hourly patterns nevertheless tend to follow a similar trend the differences tend to increase during high traffic days (Monday - Friday) and during the peak hours (11:00 - 17:00 Hrs). Due to the enormous traffic volume and chaotic access behavior, the prediction of the user access patterns becomes more difficult and complex.

Previous work presented approaches for discovering and tracking evolving user profiles. It also describes how the discovered user profiles can be enriched with explicit information need that is inferred from search queries extracted from Web log data. Profiles are also enriched with other domain-specific information facets that give a panoramic view of the discovered mass usage modes. An objective validation strategy is also used to assess the quality of the mined profiles, in particular their adaptability in the face of evolving user behavior. However the previous work concentrated only on user profiling at the application level data but not associating it to the web server. The user profile maintained by the web server enriches the user's session of authenticity at different spatial entities. The previous work used conventional web log profile analyzers weakened at the linkage of web user profiling to its server.

3) Cluster based Web Usage Knowledge Discovery Framework

The rapid e-commerce growth has made both business community and customers face a new situation. Due to intense competition on the one hand and the customer's option to choose from several alternatives, the business community has realized the necessity of intelligent marketing strategies and relationship management. Web usage mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the Web. Web usage mining has become very critical for effective Web site management, creating adaptive Web sites, business and support services, personalization, network traffic flow analysis and so on.

The proposed cluster based framework presents the important concepts of Web usage mining and its various practical applications. Further a novel approach called Web usage miner is presented. Web Usage Miner could optimize the concurrent architecture of a fuzzy clustering algorithm (to discover web data clusters) and a fuzzy inference system to analyze the Web site visitor trends. A hybrid evolutionary fuzzy clustering algorithm is proposed to optimally segregate similar user interests. Proposed approach is compared with hierarchical patterns (to discover patterns) and several function approximation techniques.

The Miner hybrid framework optimizes a fuzzy clustering algorithm using an evolutionary algorithm. The raw data from the log files are cleaned and pre-processed and a fuzzy C means algorithm is used to identify the number of clusters. The developed clusters of data are fed to a fuzzy inference system to analyze the trend patterns. The if-then rule structures are learned using an iterative learning procedure by an evolutionary algorithm and the rule parameters are fine tuned using a back propagation algorithm. The optimization of clustering algorithm progresses at a faster time scale in an environment decided by the inference method and the problem environment.

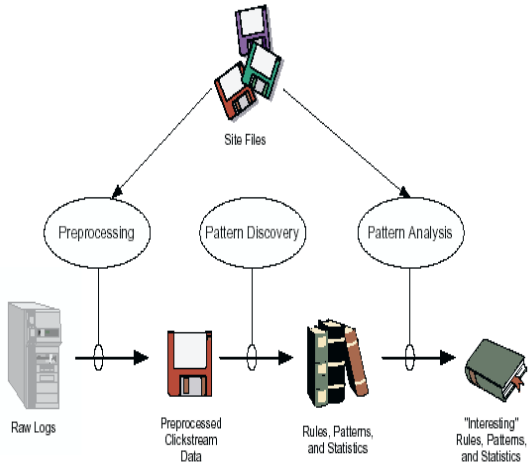


Figure 2: High Level Web Usage Mining Process

3.1 Optimization of Fuzzy Clustering Algorithm

Usually a number of cluster centers are randomly initialized and the FCM algorithm provides an iterative approach to approximate the minimum of the objective function starting from a given position and leads to any of its local minima. No guarantee ensures that FCM converges to an optimum solution (can be trapped by local extrema in the process of optimizing the clustering criterion). The performance is very sensitive to initialization of the cluster centers. An evolutionary algorithm is used to decide the optimal number of clusters and their cluster centers. The algorithm is initialized by constraining the initial values to be within the space defined by the vectors to be clustered. In the Miner approach, the fuzzy clustering algorithm is optimized jointly with the trend analysis algorithm (fuzzy inference system) in a single global search.

3.2 Expectation Maximization

Expectation maximization (EM) is used for clustering in the context of mixture models. This method estimates missing parameters of probabilistic models. Generally, this is an optimization approach, which had given some initial approximation of the cluster parameters, iteratively performs two steps, i.e., the expectation step computes the values expected for the cluster probabilities, and second, the maximization step computes the distribution parameters and their likelihood given the data. It iterates until the parameters being optimized reach a fix point or until the log-likelihood function, which measures the quality of clustering, reaches its maximum. To simplify the discussion we first briefly describe the EM algorithm.

The algorithm is similar to the Fuzzy C-means procedure in that a set of parameters are re-computed until a desired convergence value is achieved. The parameters are re-computed until a desired convergence value is achieved. The finite mixtures model assumes all attributes to be independent random variables. A mixture is a set of *N* probability distributions where each distribution represents a cluster. An individual instance is assigned a probability that it would have a certain set of attribute values given it was a member of a specific cluster. In the simplest case *N=2*, the probability distributes are assumed to be normal and data instances consist of a single real-valued attribute. Using the scenario, the job of the algorithm is to determine the value of five parameters are the mean and standard deviation for cluster 1, the mean and standard deviation for cluster 2 and the sampling probability *P* for cluster 1 (the probability for cluster 2 is *1-P*)

Algorithm Procedure

a. Guess initial values for the five parameters.

b. Use the probability density function for a normal distribution to compute the cluster probability for each instance. In the case of a single independent variable with mean μ and standard deviation $\acute{o}$, the formula is:

$$f(x) = \frac{1}{(\sqrt{2\pi}\sigma)e^{\frac{-(x-\mu)^2}{2\sigma^2}}}$$

In the two-cluster case, we will have the two probability distribution formulas each having differing mean and standard deviation values.

c. Use the probability scores to re-estimate the five parameters.

d. Return to Step b.

The algorithm terminates when a formula that measures cluster quality no longer shows significant increases. One measure of cluster quality is the likelihood that the data came from the dataset determined by the clustering. The likelihood computation is simply the multiplication of the sum of the probabilities for each of the instances.

4. Experimental Evaluation of EM and CFuzzy Cluster for Web usage Mining

Measuring the quality of the EM and CFuzzy clustering in navigation patterns mining systems needs to characterize the quality of the results obtained. The experimental evaluation was conduced using UCI repository data sets of Zoo Data Set. The data is in the original arff format used by Weka tool. The characteristics of the dataset used are given in the Table 1. Expectation Maximization Clustering Algorithm and Cfuzzy means algorithm are used for User Modeling in Web Usage Mining System.

Table 1: Dataset used in the experiments

Data set	Size (Mb)	Record- Instances
Zoo Data Set	12	1010

All evaluation tests were run on a dual processor Intel CPU 2.5 GHz Pentium Core 2 Duo

with 4GBytes of RAM, operating system Windows XP. Our implementations run on Weka tool, a data mining software for evaluation part of the system. In this study, there are two steps of data converting before applying EM clustering algorithm. There are around 800 URLs in DePaul dataset. Assigning each URL address in the session to sequential numeric values is the first step. It is impossible to assign 800 attributes to Weka so for reducing the number of attribute, each eight sequence of attributes is assigned to one attribute based on EM algorithm. Table 2 shows some basic statistics on user and sessions after cleaning, filtering and session the Zoo dataset.

Table 2: Dataset for Clustering (EM and CFUZZY)

No of Users	Size (Mb)	No of Sessions	No of Repeat Users
125	10	1028	234

5. Result and Discussion on Clusters effect on web usage patterns

EM algorithm is used in statistics for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables. The process of the algorithm repeats until likelihood is stable. Table 2 shows the cluster detection rate of the EM algorithm on the Zoo data set. The cluster detection rate is measured in terms of True positive, which is indicated in the Table 3, as the number of data record-instances increases, true positive rate of cluster object gets higher. The experiment was accomplished by maximum 114 iterations. It is complex to define the number of clusters in the initial cluster formation. In our experiment, we tried several times to tune the cluster size with other parameters to get higher true positive rate of clustering. Finally cluster the user's navigation patterns into 20 groups. Meanwhile, that the EM algorithm will get a local optimization after 26 iterations.

Table 3: Performance of EM with True Positive rate against number of record instances

No of Record Instance	Cluster True Positive Rate
1010	0.9
840	0.713
286	0.217

Table 4 depicts the performance of EM for true negative rate of cluster formation of the Zoo data set. For instance the percentage of the largest cluster is 34, while the experiment creates 20 clusters. With the increased record size, False Positive of cluster object formation also raises. However the false positive rate is comparatively very small compared to that of total number of cluster object. This in turn reduces the impurity in cluster formation. The performance graph of EM clustering on zoo data set is shown in Figure 3. Percentage of maximum cluster in clusters set will be decreased if the number of the cluster object purity is increased for every iteration.

Table 4: True Negative rate of cluster objects in web usage mining

No of Record Instances	Cluster False Positive Rate
1011	**841**

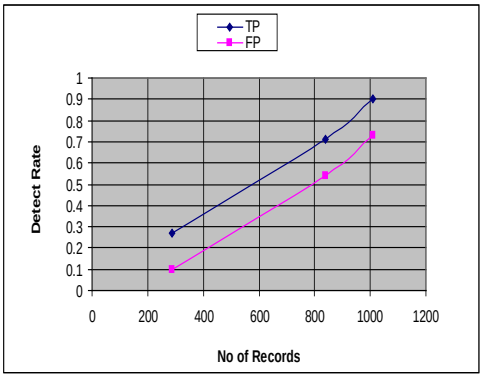


Figure 3: Performance of EM on Zoo Data set with TP and FP

The precision of the cluster formation for Zoo data set with CFuzzy cluster and EM algorithm are depicted in the Table 5. The performance of precision is measured with respect to number of instances in the data set to form the cluster with all the 10 attributes. The precision of EM is higher than the CFuzzy cluster model which is shown in Table 2 and Figure 4 indicates that the precision rate even increases for higher number of record instances of the data set.

Table 5: Performance of Cluster Precision on EM and CFuzzy algorithm

No of Record Instances	EM Precision	CFuzzy Precision
1012	0.745	0.713
842	0.682	0.645
286	0.535	0.526

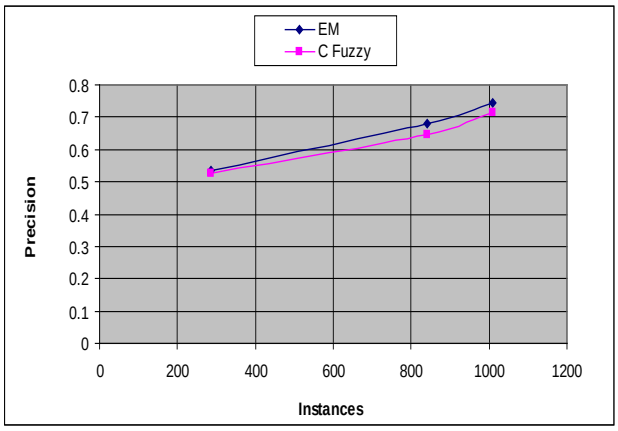


Figure 4: Comparative performance of Cluster Precision on EM against CFuzzy

Visit-coherence is utilized to evaluate the quality of the clusters (navigation pattern) produced by the EM clustering algorithm. In addition visit-coherence quantifies a session intrinsic coherence. As in the page gather system, the basic assumption here is that the coherence hypotheses holds for every session. To evaluate the visit-coherence, split dataset using 10 cross validation. The percentage of user usage visit coherence precision in the EM is approximately 11% higher than C Fuzzy means clustering algorithm for difference values of attributes.

6. Conclusion

The web usage mining framework presented in this work evaluates the performance of expectation-maximization (EM) and CFuzzy means cluster algorithms. The proposed Miner framework is an initial effort to patch up some of the weaknesses of the conventional web log file analyzers. The experimental results of EM represent that by decreasing the number of clusters, the log likelihood converges toward lower values and probability of the largest cluster will be decreased while the number of the clusters increases in each web usage pattern. The experimentation on the K-means clustering is also conducted. The results indicate the EM approach can improve accuracy of clustering to 11 more. By linking the Web logs with cookies and forms, it is further possible to analyze the visitor behavior and profiles which could help an e-commerce site to address several business questions. The further scope can be made in the direction of applying some classification methods for request. This can be used in web usage mining-based prediction systems.

7. References

[1] A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites**,** " Nasraoui, O. Soliman, M. Saka, E. Badia, A. Germain, R.", Knowledge and Data Engineering, IEEE Transactions on Feb. 2008

[2] Abraham, A (2001). Neuro-fuzzy systems: State-of-the- art modeling techniques, connectionist models of neurons, learning processes, and artificial intelligence. In Lecture Notes in Computer Science 2084, J Mira and A Prieto (eds.), Germany, Spain: Springer-Verlag, pp. 269-276.

[3] Bocca, M Jarke and C Zaniolo Analog Website Tracker (2003). (http://www.analog.cx/) [3 October 2003].

[4] Chakrabarti, S (2003). Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann Publishers.

[5] Chang, G, MJ Healey, JAM McHugh and JTL Wang (2001) Web Mining, Mining the World Wide Web Chapter 7, pp. 93-104. Kluwer Academic Publishers

[6] Chen, PM and FC Kuo (2000). An information retrieval system based on an user profile. The Journal of Systems and Software, 54, 38.

[7] Chi, EH, A Rosien and J Heer (2002). LumberJack: Intelligent discovery and analysis of web user traffic composition. In Proc. of ACM-SIGKDD Workshop on Web Mining for Usage Patterns and User Pro_les. Canada: ACM Press.

[8] Cho, YH, JK Kim and SH Kim (2003). A personalized recommender system based on web usage mining and decision tree induction, Expert Systems with Applications, 23(3), 329 - 342.

[9] Coenen, F, G Swinnen, K Vanhoof and G Wets (2000). A framework for self adaptive websites: Tactical versus strategic changes. In Proc. of the Workshop on Web mining for E-commerce: Challenges and Opportunities (KDD'00), pp. 75-80.

[10] Hay, B, G Wets and K Vanhoof (2003) Segmentation of visiting patterns on web sites using a sequence alignment method, Journal of Retailing and Consumer Services, 10(3), pp. 145-153

[11] Heer, J and EH Chi (2001) Identification of web user traffic composition using multi-modal clustering and information scent, Workshop on Web Mining, SIAM Conference on Data Mining, pp. 51 - 58.

[12] Heinrichs, JH and JS Lim (2003) Integrating web-based data mining tools with business models for knowledge management, Decision Support Systems, 35(1), pp. 103- 112.

K.Poongothai received the M.Sc (IT). Degree in Information Technology from M.Kumarasamy College of Engineering,Karur in 2006 respectively. Presently she is working in Selvam College of Technology, Namakkal, and Tamilnadu, India as Assistant Professor in Department of Information Technology

M.Parimala received the MCA. Degree in Computer Application from Mother Theresa Women's University, Kodaikanal in 2005 respectively. Presently she is working in M.Kumarasamy College of Engineering,Karur Tamilnadu, India as Lecturer in Department of Computer Application

Dr.S.Sathyabama received the M.Sc.in Avinashilingam Deemed University, Coimbatore in 1997, M.Phil in Bharathiar University, Coimbatore in 2001 and Ph.D. degree in Periyar University in 2007, Salem. She worked as Lecturer from 1997 to 2001 in karuppannan Mariappan College, Muthur. , she worked as a Professor in the Department of Master of Computer Application from 2001 to 2011 at K.S.Rangasamy College of Technology and presently she is working as Assistant Professor of Computer Science, Thiruvalluvar Govt Arts and Science College, Rasipuram.

Multi databases in Health Care Networks

Nadir K.Salih Tianyi Zang Mingrui Sun
School of Computer Science and Engineering, Harbin Institute of Technology, China

Abstract

E-Health is a relatively recent term for healthcare practice supported by electronic processes and communication, dating back to at least 1999. E-Health is greatly impacting on information distribution and availability within the health services, hospitals and to the public. E-health was introduced as the death of telemedicine, because - in the context of a broad availability of medical information systems that can interconnect and communicate - telemedicine will no longer exist as a specific field. The same could also be said for any other traditional field in medical informatics, including information systems and electronic patient records. E-health presents itself as a common name for all such technological fields. In this paper we focuses in multi database by determined some sites and distributed it in Homogenous way. This will be followed by an illustrative example as related works. Finally, the paper concludes with general remarks and a statement of further work.

Keywords: Multi databases, Health Care, Distributed Database.

1. Introduction

The advent of the internet had a major impact on the healthcare industry in the last four decades. While the sophistication of Public Digital Assistant (PDA), wireless systems and browser based technology is at the forefront of all healthcare entities considering implementation and/or expansion of their technology, there are no limits as to how far these will go. With all major financial decisions comes bench marking for best practices, conflicts and negotiations. In health care networks computers are being used with increasing enthusiasm, although the exploitation of their capabilities still lags behind that of industry in general. The 'information technology revolution' has made a great impact on daily life in modern society, being used in familiar applications from high-street cash dispensers to children's education. Increasing exposure to computing tools and services has meant that much of the mystique surrounding the discipline is disappearing, and the next generation of medical professionals and other workers in the health sector can be expected to have a very positive approach to computing [1]. Most of today's Hospital Information Systems (HIS) are characterized by a large number of heterogeneous system components [2].

A multidatabase system consists of a collection of autonomous component database systems. Distribution of data across multiple sites is a clear trend in many emerging internet applications. One major advantage of data distribution is each site can process it's own data with some degree of autonomy and user's can be provided with a single global view of the data[3]. Through Internet, the e-health care could point out record, measure, monitor, manage, and in the end to deliver patient oriented along with condition-specific care services in real time. Internet-based e-health is capable of operating ubiquitously, at anytime for anyone. E-health has been making health care more effective, allowing patients and professionals to do the previously

impossible through the widespread information and communication technologies [4]. The telemedicine system, a currently used information system, enabled to maximize the collection, delivery, and communication of health care information, clinical messages, nursing interaction, and medical records from one location to another in e-health fields [6].Nadir k.Salih et al. [8] They have recommended an agent-Web service that has the features of both the agent technology as well as the Web services technology and is managed by an autonomic system based on multi-agent support. This can help to develop enterprise IT systems that are optimal, highly available. And building deployable solutions in the number of application domains comprising complex, distributed systems.

The remainder of the paper is structured as follows: Section two presents Objectives of computerized information systems in a health network. Section 3 describes some sites and databases in health care networks. Section 4 demonstrates Homogenous Distributed Database Systems. Section 5 Related works, Section 6. Finally, the paper concludes with general summary

2. Objectives

Three factors will greatly influence the further development of information processing in health care with in the near future: the development of the population, medical advances, and advances in informatics. Healthcare in the 21st century requires secure and effective information technology systems to meet two of its most significant challenges: improving the quality of care while also controlling the costs of care. The demands of computerized information systems in a health network with regard to hardware and software are rarely matched by industrial applications. The problems to be solved are therefore correspondingly diverse and require many innovative techniques. The objectives of such computerization are to:

- Reduce the need for, and duration of, treatment of patients by good prevention methods and early diagnoses;
- Increase the effectiveness of treatment to the extent allowed by improved information;
- Relieve professionals and other workers in care units of information processing and documentation burdens, thereby freeing them for more direct care of the patient;
- Enhance the exploitation of resources available for health care by good management and administration;
- Archive clinical data, facilitate the compilation of medical statistics and otherwise support research for the diagnosis and treatment of disease.

3. Some sites and databases in health care networks

During the data analysis phase of a study of such a typical network, six important sites or functional area types were identified as providing a realistic and representative scenario for a case study. The sites were

(1) Community care units for schools and the community;

(2) General practitioner or health centre units providing a first contact point for patients with the health care network;

(3) Casualty units within hospitals for treating accidents or other emergencies requiring urgent responses;

(4) Laboratories, usually within hospitals, to provide analyses of samples from various other units; (5) Patient records offices, for the administration of medical records within hospitals;

(6) Wards for the treatment of in-patients within hospitals.

The requirements of each of these sites were distilled from the output of the data analysis phase of the study, supplemented by data descriptions reported by other studies, and expressed as a collection of five descriptions: data objects handled, functions carried out, events triggering each function; constraints on the functions; and the distribution of the functions and the data around the network.

4.Homogenous Distributed Database Systems

A homogenous distributed database system is a network of two or more databases that reside on one or more machines [7]. Figure 1 illustrates a distributed system that connects three databases: COMMUNITY CARE DATABASE, HEALTH CENTRE DATABASE and CASUALTY. An application can simultaneously access or modify the data in several databases in a single distributed environment. For example, a single query from a COMMUNITY CARE DATABASE client on local database can retrieve joined data from the PATIENT table on the local database and the DOCTOR table on the remote HEALTH CENTRE database. For a client application, the location and platform of the databases are transparent. You can also create synonyms for remote objects in the distributed system so that users can access them with the same syntax as local objects. For example, if you are connected to database COMMUNITY CARE DATABASE yet want to access data on database HEALTH CENTRE DATABASE, creating a synonym on COMMUNITY CARE DATABASE for the remote PATIENT table allows you to issue this query:

SELECT * FROM PATIENT;

In this way, a distributed system gives the appearance of native data access. Users on COMMUNITY CARE DATABASE do not have to know that the data they access resides on remote databases.

5. Related works

Mobile multi-agent information Platform MADIP that is developed on top of JADE and allows MAs to work on behalf of health care professionals, to collect distributed users' vital sign data, and to spontaneously inform abnormal situations to associated health care professionals [4]. National Health Information Network (NHIN) This model shows the feasibility of an architecture wherein the requirements of care providers, investigators, and public health authorities are served by a distributed model that grants autonomy, protects privacy, and promotes Participation [5]. Present the design and architecture of a mobile multi-agent based information platform – MADIP – to support the intensive and distributed nature of wide-area (e.g., national or metropolitan) monitoring environment. To exemplify the proposed methodology, an e-health monitoring environment was built on top of MADIP [6].

6. The appropriateness of DDB technology for health applications

There is an interesting hierarchy or network of (distributed) databases within the distributed databases for this application. There are many (distributed) databases in the system which corresponds to individual patients, doctors and many other objects. Some of these may appear as simple entities in other distributed databases. Consider an extreme case of a patient who has some chronic condition such as asthma or

hypertension, which may be particularly persistent. He or she could quite conceivably also be found, perhaps several years later, to be suffering from some other chronic condition, for example arthritis. A patient like this could, over a decade say, accumulate a sizeable collection of details resulting from frequent consultations, tests and treatments. Now all patients' records constitute a distributed database in a structural or intensional sense. However the physical size of the record of an individual chronically ill patient could mean that it qualifies as a database in the extensional sense also.

7. Conclusions

We have looked at the objectives of computerization of health care systems, and we have concentrated up on some sites and databases in health care networks. We also provide an intuitively acceptable set of criteria to help determine if the DDB approach is appropriate for a particular application environment.

We are also working with healthcare professionals to ensure that our e-healthcare system meets their needs, and we are improving our systems based on their feedback.

8. References

[1] David Bell, Jane Grimson, Distributed Database Systems, WESLEY, 1992.

[2] R. Lenz, K. A. Kuhn**,** Intranet Meets Hospital Information Systems: The Solution to the Integration Problem? Methods of Information in Medicine Schattauer GmbH (2001).

[3] Praveen Madiraju, Rajshekhar Sunderraman, a Mobile Agent Approach for Global Database Constraint, 2004 ACM Symposium on Applied Computing.

[4] Chuan-Jun Su, Chia-Ying Wu, JADE implemented mobile multi-agent based, distributed information platform for pervasive health care monitoring, Applied Soft Computing 11 (2011) 315–325.

[5] Andrew J. Mcmurry, Clint A. Gilbert, Ben Y. Reis, Henry C. Ccueh, Md, Isaac S. Kohane, Kenneth D. Mandl. A Self-scaling, Distributed Information Architecture for Public Health, Research, and Clinical Care. J Am Med Inform Assoc. 2007;14:527–533.DOI10.1197/jamia. M2371.

[6] Chuan Jun Su, Mobile multi-agent based, distributed information platform (MADIP) for wide-area e-health monitoring. Computers in Industry 59 (2008) 55–68.

[7] Jason Durbin, Lance Ashdown, Oracle8*i* Distributed Database Systems, Release 2 (8.1.6), 1999, Oracle Corporation

[8] Nadir K .Salih, Tianyi Zang, G.K .Viju, Abdelmotalib A. Mohamed. Autonomic Management for Multi-agent Systems. International Journal of Computer Science Issues, 2011.

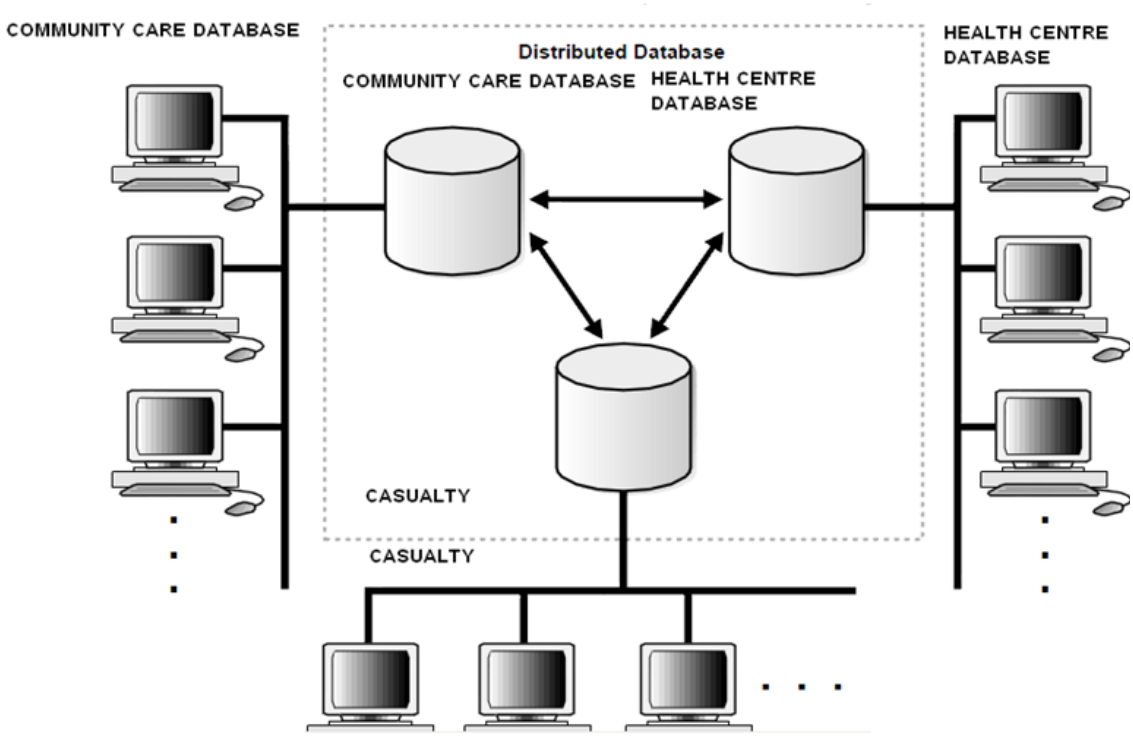


Figure 1 Homogeneous Distributed Database
(Self Creation)

Pseudonymous Privacy Preserving Buyer-Seller Watermarking Protocol

Neelesh Mehra[1], Dr. Madhu Shandilya[2]

[1] Department Of Electronics and Communication, S.A.T.I(Engg.)College, Vidisha, M.P, India.

[2] Department Of Electronics, M.A.N.I.T, Bhopal, M.P, India.

Abstract

A buyer-seller watermarking protocol utilize watermarking along with cryptography for copyright and copy protection for the seller and meanwhile it also preserve buyers rights for privacy. Up to now many secure BSW protocol has been suggested but the common problem with all of them is that in all of them is the buyer's involvement in generation of some cryptographic key or watermark or digital signature what happened if buyer is not capable or is a layman and does not understand what cryptography and watermarking means. In this paper we proposed the use of open access identification concept for this buyer has to get registered with some trusted third party which after registration provide an open access ID which is unique. This not only provide anonymity to buyer but the seller can also provide some benefit to his loyal customers. In our scheme non of the watermark or cryptographic key is generated by buyer so a layman buyer can also use it. It also enables a seller to successfully identify a malicious seller from a pirated copy, while preventing the seller from framing an innocent buyer and provide anonymity to buyer.

Keywords: **Buyer-Seller watermarking protocols; watermarking; copy protection; copyright protection**

1. Introduction

Now a days multimedia data is floating throughout the world wide web . The ease by which digital content can be stored and processed without any loss of quality resulted in illegal replication and distribution of digital content To prevent this Digital Right Management Technologies(DRM) has been developed. DRM utilize special properties of cryptography and watermarking for copyright protection of multimedia data and to prevent unauthorized use of digital content. But due to lack of implementation rule a uniform DRM system is not possible yet. Earlier research on fingerprinting schemes have been conducted by Pfitzmann etal. [1], and by Camenisch et al. [2]. The shortcoming of these schemes lies in their inefficiency. A buyer-seller watermarking protocol is one that combines encryption, digital watermarking, and other techniques to ensure rights protection for both the buyer and the seller in e-commerce. The first known buyer-seller watermark protocol was introduced by Memon et al. [3].Since the first introduction of the concept, several alternative design solutions have been proposed in [4,5,6,7].

The main feature of a buyer-seller watermarking protocol is to enable a honest seller to successfully identify a traitor from a pirated content copy, while preventing the dishonest seller from framing an innocent buyer and also preserve anonymity of buyer. A buyer-seller watermarking scheme may involves the two steps[10].

(I) A watermark is embedded by seller to identify the buyer of a digital product, such as an image.

(II) When a pirated copy is found, the seller will detect the watermark of the pirated copy and verify the buyer with the help of some trusted third party. A secure buyer-seller watermarking protocol is must consist of following properties

Traceability: A copyright violator should be able to be traced and identified. Non-framing: Nobody can accuse an honest buyer.

Non-repudiation: A guilty buyer cannot deny his responsibility for a copyright violation caused by him.

Dispute resolution: The copyright violator should be identified and adjudicated without him revealing his private information, e.g. private keys or secret watermark.

Anonymity: A buyer's identity is undisclosed until he is judged to be guilty.

Most of the proposed protocols has the above said properties, these protocols are infeasible as most of the protocols underlying the assumption that the buyer has the knowledge of cryptography and watermark. However, the buyer may have or have-not any knowledge of cryptography and watermark so the involvement of buyer must be reduced in the generation of watermark and cryptographic Key without neglecting his rights. Second, a buyer must interact with different parties many times and exchange different keys and store them this is very inconvenient to the buyer and accuse a high communication load.

Almost many of the above mentioned technical problems are solved in the scheme proposed by Alfredo Rial et al in their Privacy preserving Buyer-Seller watermarking Protocol(PBSW) based on Price Oblivious Transfer(POT) besides this some practical problem remain unsolved and need to be discussed. Firstly in this protocol Buyer has to interact many times with seller and Trusted third party

making system more complicated for buyer. Secondly buyer is anonymous for seller so seller cannot give some advantage to his loyal customers this may be against the marketing policy of many companies. Thirdly seller doesn't learn items bought by customer so he cannot planned strategically to improve his business. fourth short comes of this scheme is what happened if seller deliver wrong item in place. Lastly if buyer is corrupt and claiming the deliver items are not that item which he actually ordered there is no counterparts suggested to deal such types of practical problem. In this paper we propose a novel pseudo privacy preserving buyer-seller watermarking protocol which is capable to solve all the technical problem along with all practical problems mentioned above and overcome the drawbacks existing in the present protocols. Our protocol is easy to implement and accomplished in fewer steps causing no extra computation burden on the buyer The rest of this paper is organized as follows. In Section II, we describe our proposed watermarking protocol in detail. In Section III, we discuss the working of proposed scheme , in Section IV we analyze proposed method. Finally, in Section V, we summarize our work.

2. PROPOSED SCHEME

Our buyer-seller watermarking protocol consist of a buyer denoted by "B", a seller "S" trusted third party called as copyright Certification Authority (Bank) is a certification and registration authority which is responsible for registration of buyers(B_1,B_2…….B_n) and seller (S_1,S_2….S_n)and to embed second watermark which is invisible and used to verify the misbehavior of any buyer or seller also verify the payment condition ,So it can be any commercial bank so this trusted third party can be said as "Bank" .Finally An Judge "J", who adjudicates lawsuits against the infringement of copyright and intellectual property. The buyer-seller watermarking protocol we proposed in this section has four sub protocols: registration protocol, watermark generation protocol, watermark insertion protocol, copyright violation and dispute resolution protocol. In our protocol, we assume the following assumptions hold. (1) A public key infrastructure PKI is well developed. (2) The TTP is assumed to be trustworthy. (3) The encryption function used in the PKI, i.e. (Ek), is assumed to be a privacy homomorphism with respect to watermark insertion operation $\oplus$. By privacy homomorphism with respect to $\oplus$ we mean it has the property that $Ek(a \oplus b) = Ek(a) \oplus Ek(b)$.

Our algorithm has been accustomed of following signs

B:Buyer of certain multimedia content
S: seller of certain multimedia concept
CCA :Copyright certifying authority
sKB: private key of CCA
pKB: public key of buyer issued by CCA
TID: Transaction ID
CA: Certificate of authenticity

2.1 Registration

Registration process has two phase both are mutually explicit one is for customer and another is for seller
The registration protocol, performed between the buyers(B_1,B_2…….B_n) and the copyright Certification Authority (Bank) the registration process as follows:
Step 1. If the buyer B wants to remain anonymous during transactions, he asks Bank for an anonymous certificate and get himself registered with Bank.

Similar to above the Seller has to get registered with Certification authority and request a certificate of authenticity from the bank

Step 2. Bank now provide an open access identity which will act as pseudo-identity for the buyer and will be the identity of buyer for seller .
This registration is one time process and will be valid until any of the party involve will refuse to continue.

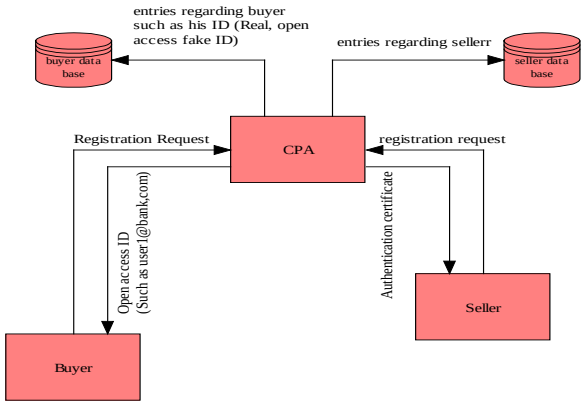


Figure 1

2.2 Initialization

The initialization of protocol starts as the buyer wish to purchase a message (m_1 to m_n), this process performed between the buyer(B) and the Certification Authority (Bank) the Initialization process as follows:

Step 1. buyer asks Bank for an anonymous certificate and sends detail of items and seller to Bank,
Step 2. Bank verify from seller about the items availability and their price and confirm the amount deposit in account of buyer .
Step 3. After confirmation Bank selects a key pair (pkB , skB) randomly, then he generates an anonymous certificate $Cert_{CCA}$(pkB) and an anonymous transaction

ID (TID) which can identity B. Bank sends certificate Cert CA(skB), skB (TID) to B and Cert CA(pkB) to S.

2.3 Watermark Generation and embedding Protocol At Seller End

Seller S generates a watermark V which can be used to identify the guilty user. This is a protocol between Buyer Seller and Trusted Third Party

Step 1. When B wants to buy a digital content X from the seller" S", B first negotiates with seller"S" to set up an agreement (ARG) which explicitly states the rights and obligations of both parties and specify the digital content X . The ARG uniquely binds this particular transaction to X and can be regarded as a purchase order.Buyer B sends his transaction ID (TID) to Seller

Step 2. Upon receiving transaction ID (TID) from B, S verify it with the TID provide by Bank, If it is valid, S generate or select (from his database) a unique watermark "VW" which is visible watermark and unique key pair(pkS , skS) for this particular transaction.

Step 3. Now this watermark is embedded by S in digital content X such that

$X' = X \oplus skS(VW)$

Step 3. Now X' is encrypted with the Private key send by buyer and send it to bank ie.S sends EpkB(X'), to bank and the public Key pkS to buyer.

2.4 Watermark Embedding Protocol At Bank End

Step 1. When Bank receives the encrypted digital product EpkB (X') from S, It decrypt it with the help of his private key afterwards bank generates or select (from his database) a watermark(IW) and a symmetric key(pKt) to insert a invisible watermark(IW) in X', buyer is anonyms of this watermarking. Seller may knows about invisible watermark but he doesn't know about what generated watermark

Such as: $X'' = EpkB[X' \oplus EpKt(IW)]$

Step 2. After that, X'' is send to B along with TID.

When B receives the encrypted watermarked X'' he decrypt and remove visible watermark by using public key pkS. The entire protocol can be summarized by seeing figure 1 and 2

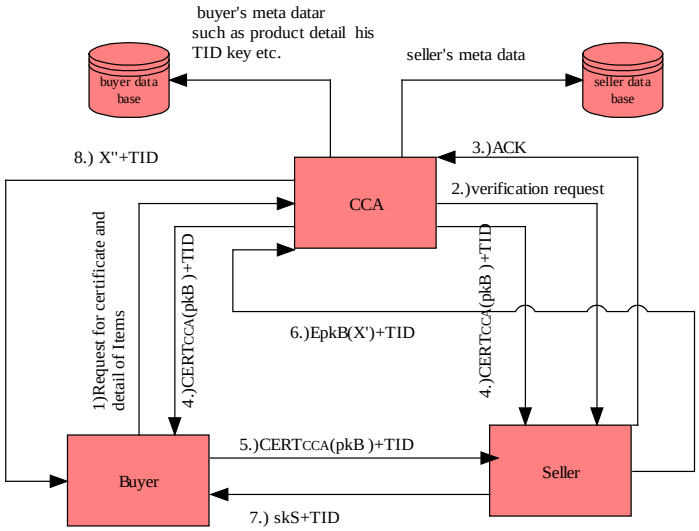


Figure 2.

Identification and Dispute Resolution:

When Seller found an unauthorized copy of content X, seller raises the matter to CCA.

1) Seller sends the pirated copy to CCA.

2) CCA extract invisible watermark and match with the database of buyer's .

3) if both invisible watermark is same then buyer is guilty and can be challenged in front of judge

4) If seller tries to frame innocent buyer then no invisible watermark is found or it will not match with the database of buyer since for every transaction CCA will choose different watermark.

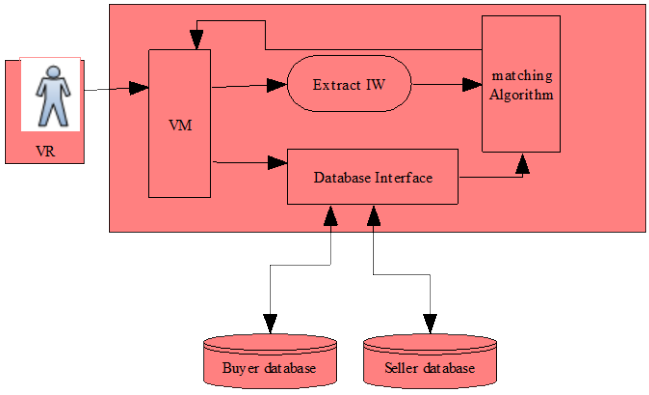


Figure 3.

3. Accountability Analysis

Ownership Right Protection : Owner of the digital work embed visible watermark with his private key into the image for owner authentication. The buyer can remove this with his public key for restriction free use of this digital work. The ownership right is now protected by the

invisible watermark which is inserted by the bank which is semi fragile in nature thus helpful in tracing malicious buyer to trace if pirated copy found. A semi-fragile watermark is sensitive to non-permitted modifications. Ideally, a semi-fragile watermark would gloss over innocent alterations on the image (for example postproduction editing, mild compression, filtering or contrast enhancement) but it should give alarm when content change occurs or in case of high compression rates. So, if a client tries to do any kind of malicious manipulations such as the addition or removal of a significant element of the image, would invalidate the image. Also an owner can prove his ownership with the help of copyright Certification Authority (Bank) . In case of any dispute copyright Certification Authority (Bank) extracts the copyright watermark from the disputed digital product and verifies the watermark and copyright information submitted by the valid requester with the information stored in the owner's database.

Client's Right Protection : If the seller wants to frame the innocent buyer he can sell the digital content with his own visible watermark but it cannot insert the invisible watermark, since the bank issue the certificate for the deal he cannot cheat buy supplying the wrong digital content. In case of any dispute, the right verification of a valid-requester (a client in this case) is done by the verification module in the CPA. CPA uses its verification module to extract the client certificate and watermark for any dispute.

Pirate Root Identification : The Pirate root identification is provided by the CPA. When a suspicious copy of the digital image is submitted to CPA, the CPA uses its verification module to extract invisible watermark using watermark extraction algorithm and match it with client database

Anonymity problem : The anonymity of the buyer can be retained during the transaction unless the buyer is judged by ARB to be guilty of piracy. In our proposed protocol, the dispute resolution protocol can be carried out well without the buyer's participation, so the buyer's privacy right is well protected.

Buyer's participation in the dispute resolution problem : Buyer's participation is not required in dispute resolution protocol with the assistance of Trusted Third Party(TTP) It can prevent malicious seller from annoying innocent buyer by repeatedly enforcing the buyer to participate in the dispute resolution.

4. CONCLUSION

In this paper we try to build a Buyer-Seller watermarking protocol which is capable to solve the common problems. Since none of the cryptographic key and watermark is generated or embedded by buyer it reduces computational cost at buyer end and make it easy to use by a laymen buyer also. Since buyer is using open access ID issued by CCA thus remain anonymous and since this ID is unique so seller can give some reward to it's loyal customer. No need of buyer's involvements in any kind of dispute until he is found guilty. Lastly the number of times of buyer's interaction with seller is less then other protocol.

References

[1] B. Pfitzmann and A.-R. Sadeghi. Anonymous fingerprinting with direct non-repudiation. In Advances in Cryptology ASIACRYPT '00, LNCS 1976, pages 401–414. Springer-Verlag, 2000.

[2] J. Camenisch. Efficient anonymous fingerprinting with group signatures. In ASIACRYPT, LNCS 1976, pages 415– 428. Springer-Verlag, 2000.

[3] N. D. Memon and P. W. Wong. A buyer-seller watermarking protocol. IEEE Transactions on Image Processing, 10(4):643–649, 2001.

[4] J.-H. P. Jae-Gwi Choi, Kouichi Sakurai. Does it need trusted third party? design of buyer-seller watermarking protocol without trusted third party. In Applied Cryptography and Network Security, LNCS 2846, pages 265–279, 2003.

[5] B.-M. Goi, R. C.-W. Phan, Y. Yang, F. Bao, R. H. Deng, and M. U. Siddiqi. Cryptanalysis of two anonymous buyer seller watermarking protocols and an improvement for true anonymity. In Applied Cryptography and Network Security,
LNCS 2587, pages 369–382, 2004.

[6] C.-L. Lei, P.-L. Yu, P.-L. Tsai, and M.-H. Chan. An efficient and anonymous buyer-seller watermarking protocol. IEEE Transactions on Image Processing, 13(12):1618–1626, 2004.

[7] J. Zhang, W. Kou, and K. Fan. Secure buyer-seller watermarking protocol. In IEE Proceedings Information Security, volume 153, pages 15–18, March 2006.

[8] M.-H. Shao. A privacy-preserving buyer-seller watermarking protocol with semi-trust third party. In Trust, Privacy and Security in Digital Business, LNCS 4657, pages 44–53, August 2007.

[9] Mina Deng, Bart Preneel,"On secure and anonymus buyer-seller watermarking protocol",in proceeding of IEEE International Conference ICIW,pages 524-529,June2008

[10] Defa Hu, Qialiang Li," A secure and practical buyer-seller watermarking protocol",in proceeding of IEEE International Conference MINES,pages 105-108,Nov.2009

[11]Huang Daren et al," A DWT Based Image Watermarking Algorithm, in proceeding of IEEE International Conference on Multimedia and Expo,pages 429-432,Aug..2001

[12] S.C.Ramesh ,M.Mohamad Ismail Majeed," Implementation of a visible watermarking in a secure still digital camera using VLSI design",in proceeding of IEEE

International Conference AFRICON,pages 798-801,Sept.2009

Comparison of Routing Protocols to Assess Network Lifetime of WSN

Owais Ahmed[1] Ahthsham Sajid [2] and Mirza Aamir Mehmood [3]

Department of Computer Science
Balochistan University of Information Technology, Engineering and Management Sciences
Quetta, Pakistan

Abstract

Rapid pace of improving technology in Wireless Sensor Networks (WSN) made it possible to manufacture low power, multifunctional sensor nodes. WSN is the set of small power energy confined sensor nodes which can be deployed in unapproachable domains. In WNS biggest constraint is to employ an efficient power consumption scheme. Different protocols were described for WSN out of which the research has been done on hierarchical (clustering) protocols to find out longer network lifetime. Low Energy Adaptive Clustering Hierarchy (LEACH), Power Efficient GAthering in Sensor Information System (PEGASIS) and Virtual Grid Array (VGA) protocols were analyzed for network lifetime by changing the sensing range of sensor nodes and increasing the network size. The sensing ranges used are 8m and 12m for 60, 90 and 120 number of nodes. The results found that PEGASIS had the consistency in network lifetime and it also supports large networks. While LEACH is more suitable for networks having less than hundred number of nodes.

Keywords: WSN, LEACH, PEGASIS, VGA, Protocol Comparison, Network Lifetime

1. Introduction

The enhancements in the technology lead the wireless communication and electronics to manufacture low power, multifunctional sensor nodes [1]. The typical architecture of wireless sensor node comprises of power source, transceiver, micro-controller, external memory, analog digital converter and sensors. The Wireless Sensor Network is the set of small power energy confined sensor nodes [2], being used widely for different applications like military [1], environmental, medical, home [6], location and movement finding and industrial [7]. The sensor nodes in wireless sensor network communicate via radio waves with other nodes as well as with base station [4]. The deployment of wireless sensor networks nodes is preferably random in most of the cases or in unapproachable places with remote monitoring. Further, sensor nodes may be equipped with the facilities of data aggregation (Data aggregation is the process of combining distributed data into high quality information) and fusion (Data fusion is a method in which different types of data from several sensors, are integrated to increase efficiency or accuracy) which provide the support to transmit partial processed data instead of raw data. On the other hand, WSN has to cope with several bottlenecks like power consumption [1], computation, communication and unreliable readings [5]. Among the mentioned constraints, power consumption requires more attention to prolong the network life span of the wireless sensor network. Thus for WSN, the routing protocols must have the capability to self-organize [1].

The routing protocols for WSN are classified into various categories shown in fig. 1 [8], out of which hierarchical routing category is selected to analyze the impact on the lifetime of the network. Because hierarchical routing protocols utilize the resources in efficient and optimized ways. In this paper three hierarchical WSN routing protocols are selected based on level of scalability and localization (the determination of the geographical locations of sensors). The goal of the paper is to analyze the impact of protocols on network lifetime on the basis of the network size with small network field.

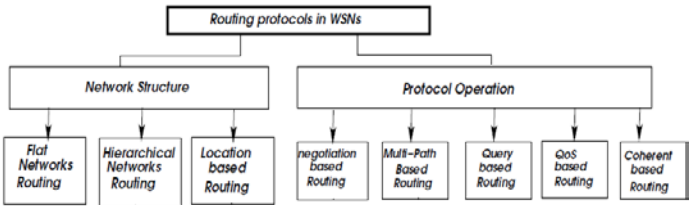


Figure1. Classification of WSN Routing Protocols [8]

may be for sensing events in the vicinity. For routing in location based networks every node is identified by its location. The

distance between the nodes can be determined by the strength of incoming signals. The other method to locate the nodes can be the implementation of GPS (Global Positioning System). The negotiation based routing protocols devastate the redundant transmission to the next sensor or BS by accompanying a series of negotiation messages. In multi-path based routing, more than one path is established between source and destination. If the primary path terminates the alternative path will be selected. In query based routing, the destination nodes broadcast a request for sensed data and the nodes having the specified data related to the query transmit back to the node. In QoS-based routing, a balance is maintained between power depletion and data quality: delay, energy, bandwidth, etc. during sending data to the sink. In last, the coherent based routing employs the minimum processing (time stamping, duplicate suppression, etc.) before sending data to the aggregators.

2. Selected Protocols

The selected three protocols are LEACH (Low-Energy Adaptive Clustering Hierarchy), PEGASIS (Power-Efficient GAthering in Sensor Information Systems) and VGA (Virtual Grid Architecture).

2.1. LEACH

All nodes are organized as set of clusters. Each cluster has a cluster head to communicate with Base Station. The cluster heads are selected on rotation bases to balance the load of energy in the way that most of the nodes get small distances to transmit and only cluster heads are responsible for long transmission to the BS. Besides, LEACH allows data fusion and aggregation in order to minimize the amount of data to be transmitted. Because for energy concerns local computations require less energy than transmitting signals to BS [9] [10].

Each round of LEACH protocol is composed of 'setup phase' and 'steady-state phase'. In setup phase the cluster heads broadcasts an advertisement message to all the nodes to elect cluster head. And the cluster heads are elected depending upon the predefined specified percentage of cluster heads and how many times the node has been elected as cluster head. On receiving the advertisement messages from cluster heads, the non-cluster head nodes decides to which cluster head it will belong depending upon the energy required for transmission to the cluster head. Thus nodes become the members of the cluster requiring low energy transmission for the cluster head [9] [8]. Each non-cluster head node sends a message to the cluster head declaring that it belongs to its cluster after the selection of that cluster head. The cluster head then generates a TDMA schedule for communication with the nodes within its cluster. In steady state phase non-cluster head nodes transmit their data only when their allocated time slots arrive. The radio of each non-cluster head node is kept off all the time except when it is ready to transmit data to BS (when its time slot arrive), reducing the battery power consumption. Furthermore, as the cluster head receives all the data from all the nodes it aggregates and fuses the data to minimize the amount of long distanced transmission with the base station. Thus again reducing the energy consumption. When a node decides to become a cluster head, it also chooses a CDMA code from the available list of spreading codes and informs all the non-cluster head nodes within its cluster about the details of the chosen code. The reason for this is that, the radio transmission of a node with cluster head in a cluster usually affects the transmission in the neighboring clusters. By CDMA, the cluster head filters the received signal using the specific spreading code [9] [8].

2.2. PEGASIS

It is chain based architecture in which transmission occurs in such a way that node send and receive data only from the closest neighbor. PEGASIS allows data to be fused but doesn't support data aggregation. On receiving data, node fuses with its own data and forwards to the next node. Node acting as a chain leader is responsible to communicate with the BS. In each round the chain leader is changed to balance the remaining energy of the network, which in turns results the longer network lifetime of sensor nodes. The chain formation is held by greedy approach which works quite well [12] [8] [10]. Each node selects its nearest node as a neighbor, starting from the farthest node from Base Station. The closest node is assessed by the signal strength from all the nodes in its surroundings. Like LEACH, PEGASIS also choose the chain head (cluster node) randomly for routing to the BS. Each node is chosen as chain head once in every N number of rounds (where N= no. of nodes) [8].

Once the chain is created and chain head is chosen, the chain head of the current round initiates a token for the end node of the chain to start the transmission. Each node except the farthest node of the chain fuses its data with the received data and sends a single packet to the next neighbor. The chain head communicates with BS after receiving the data from each side of the chain and fusing its own data. If a node dies in the chain, the chain will be reconstructed again to bypass the dead node [8].

2.3. VGA

It is a GPS-free technique to split the network topology into logically symmetrical, side by side, equal and overlapping frames (grids) [11] [13] [8]. And the transmission is occurred grid by grid [14]. VGA provides the capability to aggregate the data and in-network processing to increase the life span of the network. Data aggregation is done in two steps i.e. first at local level (in grid) and then globally. The nodes that are responsible to aggregate data locally are 'local heads' (grid heads) and the nodes 'global heads' have to aggregate data received from local heads [8] [14]. After the formation of logical grids, election is started in each grid to decide for the local head of the grid based on node the energy and how many times it has been selected as local head. And then the global heads are also selected randomly from the selected local heads. Several local heads may connect to the global head [8] [11]. The local heads are allowed to communicate vertically and horizontally only. Each node within the grid that has the required data will send its data to the local head. Then the local head will aggregate the data and send it to its associated global head that will also aggregate the data again and send it to the BS via other global heads [13]. If a local head or global head dies, a new local/global head is selected after the election [14].

3. Related Work

In [5] LEACH, PEGASIS and VGA routing protocols were compared for network lifetime on the basis of transmission range. The experiments showed that the by increasing the transmission range PEGASIS increased the total network life span. LEACH showed longer network lifetime than VGA because of the early death of the sensor nodes. While VGA affects the network connectivity badly but reposts more power when transmission range was increased.

In [15] AODV and DSR were evaluated for performance using normalized routing overhead, PDR (packet delivery ratio) and end to end delay as metrics having the variables pause time and no. of sources. It was concluded that DSR attained the better edges than AODV pertaining to overhead and PDR in restricted conditions. The results also showed that end to end delay for DSR is greater than AODV. Finally implementing large value of pause time enhanced the performance of DSR and AODV protocols.

In [16] the rate of mobilization, pace of location change and routing overhead were used as matrices. The derived results expatiated that DSR performed well for all rates of mobilization and pace to change location even of being accountable to increase the source routing overhead. AODV also achieved the same level of performance in addition to decrease the source routing overhead but is much more costly for high rates of mobilization than DSR. At last, DSDV could not attain the performance comparable to AODV and DSR when the power for transmission is amplified. Nevertheless the routing load of AODV is also boosted.

In [17] TinyAODV (AODV version for WSN), MultiHopRouter (algorithm for OSPF), GF-RSSI and GF are the 4 protocols that were analyzed claiming PDR and the energy consumption as the metrics. The results were evaluated which described that the GF-RSSI generated high packet delivery ratio and reduced power utilization. The performance of the metrics for MultiHopRouter was disgraced as the data rates were increased. Finally, high power consumption was examined in the case of TinyAODV.

In [19] the two different mobility models; constrained mobility (CM) and attenuation factor (AF) were the constituents of the experiment. The simulations for the real environment were based on the 3 matrices; packet delivery latency, packet delivery ratio and routing overhead. The results for indoor environment showed that the simplicity in the mobility models do not cause any change for DSDV contrary to DSR. Further different protocols do not produce pure results in certain scenarios as for the network of 20 nodes; mobility models did not affect the performance of DSR. But for 50 nodes network, mobility models had the impacts on the performance of DSR. Finally, the author suggested that for reliable simulations more work must be done on realistic models.

In this paper LEACH, PEGASIS and VGA routing protocols were compared for network lifetime on the basis of network size in small network field contrary to [5] in which network lifetime was assessed by changing transmission range. Further AODV, DSDV and DSR work in ad-hoc networks while TinyAODV is the version of AODV for WSN and is not hierarchical as for transmission it broadcasts the packets.

4. Simulation Scenarios

For all scenarios the common parameters for simulation include the standard values i.e. initial energy of 0.5 joules for each node, first energy model, transmission range of 15 m (except for PEGASIS which is 56.56854 m), random topology of sensor nodes deployment, network bandwidth of 5000 b/s with transmission speed of 100 b/s, data packet size of 2000 b with data processing delay of 0.1 ms, control packet size of 248 b and sensing cycle of 1 sec. The other common parameters are network field of 40 × 40 m2 with 1 BS located at (140, 25), homogenous (having same level of initial energy for all nodes) type for

temperature detection, Besides, each protocol was tested for 60, 90 and 120 nodes with both 8 m and 12 m sensing ranges to evaluate the performance of protocols for small (60 nodes) and larger (90 and 120 nodes) networks. The sensing range of 12 m (standard value is 8 m) is selected to analyze the impact of protocols on energy by increasing the sensing area of the sensors. The metric for which the simulation is conducted is 'Loss of Network Connectivity or Network Lifetime'.

5. Results

The cumulative results of the whole experiment are shown figure 2. The figure describes the number of rounds for each protocol against the sensing range (i.e. 12 m and 8 m).

Table 1: STANDARD Deviation of each protocol

No. of Nodes	**60**		**90**		**120**	
Sensing Range	***12 m***	***8 m***	***12m***	***8m***	***12m***	***8m***
LEACH	37.86	28.44	74.45	60.50	20.82	32.69
PEGASIS	18.96	13.11	27.47	9.59	14.81	29.88
VGA	30.57	33.60	15.73	120.28	99.41	26.73

The above table I. depicts the standard deviation of 5 repetitions for network lifetime of each protocol in each scenario.

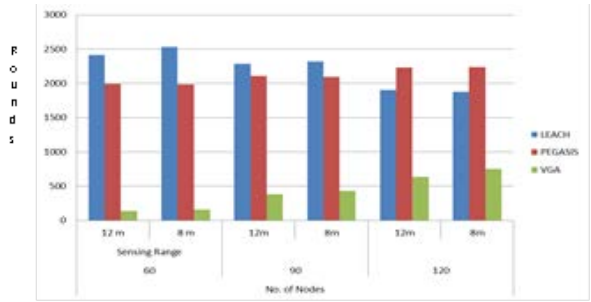


Figure 2. : Network Lifetime for 60, 90 and 120 Nodes

It is clearly evident from figure 2 that for 60 nodes, total network lifetime of LEACH is much longer than the other protocols. This is because LEACH architecture provides the support to reduce the transmission cost for less number of nodes. On the other hand the total network lifetime of PEGASIS is much higher than VGA. The overheads for grid establishment and the selection of local and global aggregators in VGA are higher. This results in the high energy consumption leading to the shorter network lifetime of VGA. In case of the 90 nodes it can be observed that LEACH protocol still remains at the top for higher total network lifetime for both sensing ranges. While in the case of PEGASIS the total network lifetime has been improved and approaching to the total network lifetime of LEACH. For VGA's total network lifetime is again much shorter among the three WSN hierarchical routing protocols due to the increase in overhead. Finally, for 120 nodes PEGASIS achieves the highest performance among the three protocols. The simulation shows that LEACH works well for less than 100 nodes. As the number of nodes increases, the overhead of cluster formation, cluster head selection and scheduling in each round also increase substantially affecting the network lifetime. While on the other hand PEGASIS has the ability to support large networks with longer network life. The reason for this is that PEGASIS creates chain only in the beginning or when a node dies, contrary to LEACH. The VGA is still far behind pertaining to network lifetime.

6. Conclusion

The growing pace of technology has opened the way to monitor and control the environment where the human interaction was not easy or even impossible. Wireless Sensor Networks usually do not require any physical interaction for maintenance and controlling that is why sensor networks are getting higher demand for future system monitoring and controlling. For WSN the main constraint is the efficient power consumption which is the great obstacle for performing tasks continuously. For this reason several techniques and architectures have been described, out of which one is the use of efficient protocol which can reduce the power consumption during communication to prolong the network life time. That is because of the fact that the communication is much expensive in terms of energy consumption as compared to the processing.

In this research, three Wireless Sensor Networks protocols (LEACH, PEGASIS and VGA) are compared to find out the performance pertaining to network life time. All the three protocols are classified as hierarchical which makes them to operate more efficiently than previous techniques like flooding. Though wireless sensor networks do not have static topologies and infrastructures but the support for dynamic hierarchy lets these protocols to work longer. While being hierarchical all the protocols have different architectures due to which their performances vary.

The three hierarchical protocols are compared for network lifetime by changing the sensing range of the sensor nodes and increasing the size of network.

The clustering architecture of LEACH makes it possible to reduce the transmission by data aggregation which minimizes the number of packets to be transmitted. Experiment showed that the performance of LEACH is much superior for smaller network (i.e. less than 100 nodes) as compared to PEGASIS and VGA. PEGASIS has shown some consistency in network lifetime for all scenarios and the ability to support large networks. It is also concluded that VGA has huge overhead and the rapid energy depleted regions in the network results in increasing the transmission path and decreases network lifetime.

References

[1] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam and E. Cayirci, "Wireless sensor networks: a survey", Elsevier Science B.V. PII: S13 8 9-1 2 86 (0 1)0 03 0 2- 4, 2002

[2] "Sensor node" available at http://en.wikipedia.org/wiki/Sensor_node, Last accessed September 29, 2010.

[3] Cong Wang and Cuirong Wang, "A Concentric Data Aggregation Model in Wireless Sensor Network", PIERS Proceedings, Beijing, China, March 23-27, 2009, 436-441.

[4] Norman Dziengel, Georg Wittenburg, and Jochen Schiller, "Towards Distributed Event Detection in Wireless Sensor Networks", April 2008.

[5] Youn Yao and Johannes Gehrke, "The Cougar Approach to In-Network Query Processing in Sensor Networks", Newsletter, ACM SIGMOD Record, Volume 31 Issue 3, September 2002.

[6] Rajashree.V.Biradar, V.C .Patil , Dr. S. R. Sawant and Dr. R. R. Mudholkar "CLASSIFICATION AND COMPARISON OF ROUTING PROTOCOLS IN WIRELESS SENSOR NETWORKS", Special Issue on Ubiquitous Computing Security Systems, UbiCC Journal – Volume 4.

[7] Ian F. Akyildiz and Erich P. Stuntebeck, "Wireless underground sensor networks: Research challenges", E.P. Stuntebeck / Ad Hoc Networks 4 (2006) 669–686.

[8] Al-Karaki, J. N. and A. E. Kamal, "Routing techniques in wireless sensor networks: A survey," IEEE Trans. Wireless Communications, Vol. 11, No. 6, 6-28, 2004

[9] Wendi Rabiner Heinzelman, Anantha Chandrakasan, and Hari Balakrishnan, "Energy-Efficient Communication Protocol for Wireless Microsensor Networks", Proceeding HICSS '00 Proceedings of the 33rd Hawaii International Conference on System Sciences-Volume 8, IEEE Computer Society Washington, DC, USA ©2000

[10] Kemal Akkaya and Mohamed Younis, "A survey on routing protocols for wireless sensor networks", Elsevier B.V, 2003, 325–349.

[11] Laiali Almazaydeh, Eman Abdelfattah, Manal Al- Bzoor, and Amer Al- Rahayfeh "PERFORMANCE EVALUATION OF ROUTING PROTOCOLS IN WIRELESS SENSOR NETWORKS", International Journal of Computer Science and Information Technology, Volume 2, Number 2, April 2010, 64-73.

[12] Lindsey, S. and C. Raghavendra, "PEGASIS: Power-Efficient GAthering in Sensor Information Systems," Proceedings of IEEE Aerospace Conference, 1125-1130, Montana, USA, Mar. 2002.

[13] Jamal N. Al-Karaki and Ahmed E. Kamal, "End-to-end support for statistical quality of service in heterogeneous mobile ad hoc networks", Journal, Computer Communications archive Volume 28 Issue 18, November, 2005.

[14] Ting-Hung Chiu and Shyh-In Hwang "Efficient Fisheye State Routing Protocol using Virtual Grid in High-Density Ad-Hoc Networks", National Science Council, Taiwan, R.O.C, 2006.

[15] Yogesh Chaba, Yudhvir Singh and Manish Joon, "Simulation based Performance Analysis of On-Demand Routing Protocols in MANETs", Second International Conference on Computer Modeling and Simulation, 2010.

[16] Guntupalli Lakshmikanth, Mr. A. Gaiwak and Dr. P .D. Vyavahare, "Simulation Based Comparative Performance Analysis of Adhoc Routing Protocols", TENCON 2008 - 2008 IEEE Region 10 Conference.

[17] Nam N. Pham, Jon Youn, and Chulho Won, "A Comparison of Wireless Sensor Network Routing Protocols on an Experimental Testbed", SUTC '06 Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing - Vol 2 - Workshops - Volume 02.

[18] Jamal N. Al-Karaki and Ghada A. Al-Mashaqbeh, "SENSORIA: A New Simulation Platform for Wireless Sensor Networks", IEEE, International Conference on Sensor Technologies and Applications, 2007

[19] Amr M. Hassain, MohamedI. Youssef and Mohamed M. Zahra, "Evaluation of Ad Hoc Routing Protocols in Real Simulation Environments", IEEE, International Conference on Computer Engineering and Systems, 2006, 288 - 293

Unsupervised Graph-based Word Sense Disambiguation Using lexical relation of WordNet

Ehsan Hessami[1], Faribourz Mahmoudi[2] and Amir Hossien Jadidinejad[3]

[1] Islamic Azad University, Qazvin Branch, Qazvin, Iran

[2] Islamic Azad University, Qazvin Branch, Qazvin, Iran

[3] Islamic Azad University, Qazvin Branch, Qazvin, Iran

Abstract

Word Sense Disambiguation (WSD) is one of tasks in the Natural Language Processing that uses to identifying the sense of words in context. To select the correct sense, we can use many approach. This paper uses a tree and graph-connectivity structure for finding the correct senses. This algorithm has few parameters and does not require sense-annotated data for training. Performance evaluation on standard datasets showed it has the better accuracy than many previous graph base algorithms and decreases elapsed time.

Keywords: *word sense disambiguation, tree, Graph connectivity.*

1. Introduction

The objective of word sense disambiguation is identifying the correct sense of word. Since Human language includes many ambiguity words. WSD is one of the essential tasks in the most Natural Language Processing (NLP), including information retrieval, information extraction, question answering and machine translation. For instance, the term of *bank* has two senses: *finance* and *shore*. The correct sense of an ambiguous word can be selected based on the context where it occurs. The problem is defined as the task of automatically assigning the appropriate sense to polysemous word at given context.

The methods of word sense disambiguation can be classified in Supervised and Unsupervised. The supervised approaches have the better performance than unsupervised approaches [1,2], the supervised systems accuracy are between 60 and 70 percent and the unsupervised systems are between 45 and 60 percent [2]. But often require large amounts of training data to yield reliable results and their coverage is typically limited to the some words. Unfortunately, creating a suitable train-data which is including all the human language words and sense are too difficult, expensive and must be reiterated for new domains, Words, and sense inventories. Therefore, these approaches have many problems. As an alternative to supervised systems, knowledge-based WSD systems extract the suitable information and present in a lexical knowledge base to perform WSD, without using any further corpus evidence. The unsupervised methods can be used this lexical knowledge-based to WSD.

In the field of WSD, the unsupervised approaches are used to methods that perform sense disambiguation without need to train data. The unsupervised approaches divided in two classes: graph-based [6,11,14,15,16] and similarity-based [3,8,12]. Graph-based algorithms often have two steps. First, construct semantic graphs from words of context, and then process the graph in order to select the best sense for each of words. Similarity-based algorithms assign a sense to an ambiguous word by comparing each of its senses with those of the surrounding context, then select the sense has highest similarity. Experimental comparisons between the two algorithm types indicate that graph-based algorithms have better performance than similarity-based [5].

This paper, describes a different graph-base algorithm for unsupervised word sense disambiguation, builds a tree and finds the best Edges. Then builds a graph with the edges and uses the connectivity measure methods for extract the best sense of each word. Also uses the WordNet efficiently, performing significantly better that previously published approaches in English all-words datasets. Show that the algorithm has good results, also present some condition for receiving the better result, performance and time consuming.

The paper is organized as follows. We first describe Related work and followed by WordNet. Section 4 describes proposed algorithm. Section 5 shows the experimental setting and the main results, finally we conclude with a discussion of the conclusion and future works.

2. Related Work

In this section, briefly describe some graph-based methods for knowledge-based WSD. All the methods rely on the information represented on some lexical knowledge base, which typically is some version of WordNet, sometimes enriched with proprietary relations. The results on datasets show in Table 2.

Mihalcea [13] presented an approach that used the PageRank algorithm to identify sense which is relevant in context. Initially, builds a graph from the possible senses of words in a text and interconnects pairs of senses with meaningful relations by WordNet. Graph edges have weight. The weight of the links joining two synsets is calculated by executing Lesk's algorithm between them. Then, use the application of PageRank for selecting the best sense of each word. The PageRank computations require several alternatives through the graph to achieve the suitable ranking for sense of word.

Navigli and Velardi [4] presented the *Structural Semantic Interconnections*(SSI) algorithm, that offered method for development of lexical chain base on the encoding of a context free grammar of valid semantic interconnection patterns. To find the meaning of the words in WordNet glosses used, but can be used for English-all words, though has the weakly accuracy. Given a text sequence, first identifies ambiguity words and builds a sub graph of the WordNet lexicon which includes all the senses of words. Then, select the senses for the words which maximize the degree of connectivity of the induced sub graph.

Navigli and lapata [5] presented a method for build a graph, that had few parameters and did not require sense-annotated data for training. First, added the sense of words in a set, then for the all of sense perform a Depth-First Search (DFS) of the WordNet graph. If appear the node is a member of set, will add all the intermediate nodes and edges on the path in the set. Finally, uses the graph connectivity measures for selecting the best sense for each of words. Also present a study of graph connectivity measures for unsupervised WSD and indicated that the local measures performance is better than global measures. The best local measures are Degree and PageRank.

Sinha and Mihalcea [6] extend their previous work on unsupervised graph-based method for word sense disambiguation by using a collection of semantic similarity measures when assigning a weight to the links across synsets. Also presents and performs this system with all the measures of word semantic similarity and graph connectivity measures. Also Showed that the right combination of word similarity metrics and graph centrality algorithms can significantly outperform methods proposed in the past for this problem, therefore reduces 5–8% of error rate.

Agirre and Soroa [11] proposed a new graph-based method that uses lexical knowledge base in order to perform unsupervised word sense disambiguation. They create a sub graph of WordNet which connects the senses of the words in the input text, and then use Personalize PageRank. Performance is better than previous approaches that used PageRank in English all-words datasets. Also show that the algorithm can be easily ported to other languages with good results. The good choice of WordNet versions and appropriate relations are fundamental to the system performance.

3.WordNet

WordNet is an ontology of lexical which created and maintained at Princeton University. The WordNet lexicon contains nouns, verbs, adjectives, and adverbs. Senses of lexical have relation with together. The words that have similar sense encodes in synonym sets (henceforth synsets). Wordnet 3 is the latest version, contains approximately 155,000 words that organized in 117,000 synsets [1,4].

Relations have been organized in two sets, Lexical and semantic relations. The lexical relations are used to connect between the lexical and the semantic relations for the synsets. For example Antonymy, Pertainymy and Nominalization are lexical relations and Hypernymy, Holonymy, Similarity are semantic relations. Also provide a textual definition of the synset possibly with a set of usage examples, that's called gloss. Figure 1, shows the WordNet semantic networks of car_n^1 synset[1].

4. Proposed Method

This section, describes a proposed algorithm. Algorithm proceeds incrementally on a sentence by sentence basis. When given a sentence, is a sequence of words W = $\{w_1, w_2, w_N\}$, assumed the sentences are part-of-speech tagged, so considers content of

words only (i.e., nouns, verbs, adjectives, and adverbs).

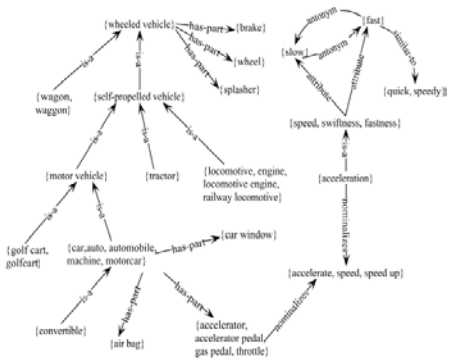


Figure 1.The WordNet semantic networks

Algorithm has two base sections, In First section, to enhance the algorithm performance, omits the stop words[1]. Then, for each of w_i, must extract the senses from WordNet that have specified part-of-speech and Tag-Count is greater than zero, S_{w_i}= $\{S^1_{w_i}, S^2_{w_i}, \ldots, S^n_{w_i},\}$. Tag-Count is frequency of this word sense measured against a text corpus. After that, add the senses in set G . G uses for graph G= (V , E). V is a set of nodes and E is a set of edges respectively, V = { $S_{w_i}|i = 1..N$} and E = Ø.

For each of S_i in G must build a tree . This tree builds from the relations of WordNet. To improve performance of the algorithm only use relevance lexicalizes and words. Furthermore, the lexical will be added within the tree when it does not appear in the previous levels of tree. Thus, the nodes of tree are lexical and the edges are lexical relations. The depths of tree is denote with *maxlevel.* The *maxlevel* is a maximum level of the tree.

In second step, search the tree to find a node that is a member of G. If it found, an edge would be added from the root's tree to specific node. Finally, use the connectivity measure method to select the best sense for each of words. Sometimes none of word senses are a member of the graph. For these words select the sense that has the highest probability (the first sense) which is the common sense.

For example, assume we have the sentence "*he drinks some milk"*. Initially omit the stop words are (*he, some*). Then, extract the senses of *Drink* and *Milk* from WordNet. *Drink* in this sentence is verb and *Milk* is noun. WordNet for *Drink* has five senses and four senses for *Milk*. But, only four senses for *Drink* and two senses for *Milk* have the Tag-Count greater than zero. Add these senses in set G. Therefore, must build the tree for all of members$G$. Figure 2 shows the tree of $drink^1_v$.

$$G = drink^1_v, \ldots, drink^4_v, milk^1_n, milk^2_n$$

After completing the tree, search in the tree for finding the nodes that are a member of G.In figure 2 shows target nodes denote with green. If target nodes found, the edge would be added in set G . Figure 3 shows the finally graph.

Now use the one kind of connectivity measure for select the best sense. Here first sense is better sense for *Drink* and *Milk*. Duo to, have the most arrival connectivity.

Algorithm 1. Propose method For Word Sense Disambiguation.

Input: Sequence W = {w_i |i = 1..N}

Extract senses
1: **for** each of words **do**
2: extract the senses that have Tag-count > 0.
3: add s_i in G.
4: **end for**

Build Tree and Graph
1: **for** each of s_i in G **do**
2: **While** level of tree <= *maxlevel* **do**
3: **for all** nodes of tree (v_i) **do**
4: **for all** the WordNet lexical relations of v_i **do**
5: **if** lexical not exist in the tree **then**
6: add the lexical in the tree.
7: **end if**
8: **end for**
9: **end for**
10: **end while**
11: **If** find the nodes are member of G **then**
12: add edge form s_i to nodes in G.
13: **end if**
14: Delete Tree.
15: **end for**

Score vertices in G
1: **for all** vertices in G **do**
2: Score(v) ← Degree Centrality(v).
3: **end for**

Sense assignment
1: **for** each of words **do**
2: sense of word ← max (Score(v)).
3: **end for**
4: **If** words don't have the sense in G **Then**
5: sense of word ← first sense.
6:**end if**

[1]http://www.webconfs.com/stop-words.php

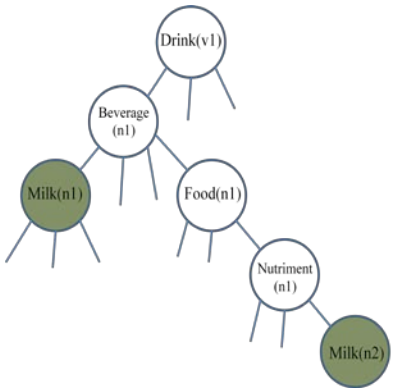


Figure2. Tree of Drink(v1)

5. Experiments And Result

In order to speed up and enhance accuracy the word sense disambiguation, we use the tree structure and prune some relations. Moreover, all paths connecting pairs of senses in WordNet were exhaustively enumerated and stored in a database. Also determine the best maximum value for depth of the tree experimentally. Run WSD algorithm on the Sensaval-3 data set using the Degree connectivity measure and the WordNet sense inventory while varying the depth length from 3 to 6. The length 3 isn't very good. The length 5 and 6 are very time consuming and their accuracy are not better than 4.

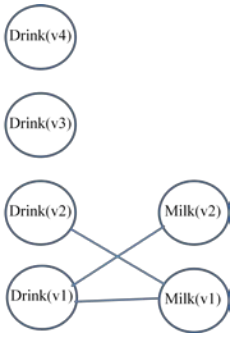


Figure 3.Graph for the sentence he drank some milk (Drink, Milk).

Therefore, we choose 4 for the depth path of the tree.
In order to select the best sense for the words in graph can use local and global measure methods. Local measures of graph connectivity determine the degree of relevance of a single vertex in a graph. But, Global connectivity measures are concerned with the structure of the graph as a whole rather than with individual nodes. Navigli in [5] indicated that local measures yield better performance than global ones, and the degree centrality that is local measure had the best result for the graph. Degree centrality is the simplest way to measure node, it is the degree of node that normalized with maximum degree [7]. This paper used the degree centrality.

5.1. Data

Evaluation and comparing the word sense disambiguation systems is very difficult, because each other use the different data set, knowledge resources and sense inventory. *Senseval*[2] (now renamed *Semeval*) is an international word sense disambiguation competition. The objective is to perform a comparative evaluation of WSD systems in several kinds of tasks, include all-words and lexical sample WSD for different languages. The Senseval workshops are the best reference to study the recent trends of WSD.
This paper evaluated the experiments on the Sensaval-2 [9] and Sensaval-3 [10] English-all words data sets. These data sets were manually annotated with the correct senses by human and use for competitions and evaluation the different systems. The sensaval-3 is difficultly to disambiguate than sensaval-2, but the Senseval-2 data set is meaningfully than the Senseval-3 data set, thus more appropriate as a test set [6]. These data-sets labeled withWordNet1.7 tags. These were normalized to WordNet 3.0 using publicly available sense mappings[3]. Table 1, shows the statistics of those data sets.

Table 1.Occurrences of noun (N), verb (V), adjective (Adj.) and adverb (Adv.) words of Wordnet 3 in Senseval 2 and Senseval 3.

Sensaval-2				Sensaval-3			
N	***V***	***Adj***	***Adv***	***N***	***V***	***Adj***	***Adv***
1136	581	457	299	951	751	364	15

5.2. Results

This section provides an evaluation the tasks that Described in the previous section. The base algorithm, uses the all relation of WordNet and all the words in a sentence (denote AT-A), Also extracts the senses from WordNet that have the Tag-Count are greater than zero. With these conditions in Senaval-3 accuracy and recall are 52.67% and Senaval-2 is 58.67%. The AT-A problem is time consuming, due to using the all words in the sentence. If omit the stop words, in the Sensaval-3 accuracy is 56.52% and recall is 44.16 %, also the Sensaval-2 accuracy and recall are 62.18% and 47.23 % respectively. This method Denotes with WT-A. When omit the stop word, the accuracy and system performance are improved, but reduced the recall.

[2] http://www.senseval.org.
[3] http://www.cse.unt.edu/~rada/downloads.html

Therefore, the stop words don't have specific sense and may add the noisy edge in the graph.
In order to reduce time and enhance the performance and accuracy, use only lexical relation for building the tree and omit some nodes and extract the senses from WordNet that have the Tag-Count are greater than zero. It's our proposed algorithm (denote WT-R). With this condition in Sensaval-3 accuracy is 63.28% and recall is 49.45% and in Sensaval-2 accuracy and recall are 65.00% and 49.41% respectively. This has very good time and accuracy, because use only lexical relation and prune some nodes and senses.
Table 2 compares the accuracy of the best graph-based method with our methods. As discussed in Sec 2. Mihalcea et al. [13] (Mih05), the method of Agirre and Soroa [11] (Agi09), the results from the work of Navigli and Lapata [7] (Nav07), the method of Navigli and Velardi [4] (SSI), the method of Navigli and Lapata [5] (Nav10) and the method of Sinha and Mihalcea [6] (Sinha07) are well-known methods in the literature.

Table 2. Comparison with related work

	Sensaval-2	**Sensaval-3**
	Accuracy	***Accuracy***
Mih05	54.2	52.2
Agi09	59.5	57.4
Nav07	n/a	52.5
SSI	n/a	60.4
Nav10	n/a	52.9
Sinha07	56.4	52.4
AT-A	58.67	52.67
WT-A	62.18	56.52
WT-R	**65**	**63.28**
FS	63.7	61.3

Whenever results were not available, due to they were not reported in the literature, an entry *n/a* exists in the respective cell. Finally, added in the comparison a simple heuristic method (*FS*) that always selects the first sense of the target word from WordNet (i.e., the most frequent) to conduct the disambiguation.
Figure 4, Show Elapsed time (in minutes) of our algorithms when applied to the Senseval-3 dataset. The proposed method has very good time, use the some relations of WordNet. Also, with omit stop word the performance of system is better than when use the all words. These times acquired by a computer with processor 2.50GHZ Core 2 Duo and 4GB RAM.

6. Conclusion And Future Work

This paper has proposed a new method for word sense disambiguation. First builds a tree for some of the senses of ambiguity words which there are in the sentence and detects the best path. Then with these paths builds a Graph and uses the connectivity measure for choosing the best sense of words. Here, we used the degree centrality, because Navigli [5] proved it's the best connectivity measure. When we are building the tree, uses some relations of WordNet to improve the accuracy and performance system together. The previous methods used the all relation and senses of word for WSD. But, we use only lexical relation and the senses that have the Tag-Count are greater than zero. With this condition our

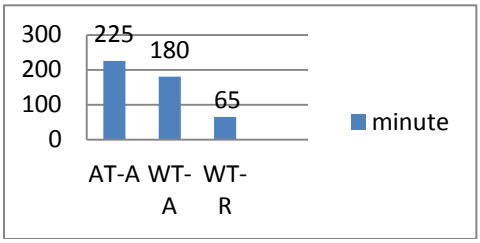


Figure 4. Elapsed time (in minutes) of the algorithm when applied to the Senseval-3 dataset

graph is less than other methods and compute is easier. Also the result is better than other graph-base and other unsupervised method. Performance our proposed method (WT-R) is greater than 20 percent better than base method (AT-A) and accuracy is 63.28 percent in sensaval-3 and 65.00 percent in sensaval-2 dataset. The algorithm can be applied easily to sense inventories and knowledge bases different from WordNet.
In the future, we are interested in applying the proposed method to weight graphs. For this purpose we can use the measures of word semantic similarity or Navigli proposed graph [6] with other conditions and calculate the probability of nodes in graph connectivity.

Refrences

[1] R.Navigli," Word sense disambiguation: A survey", *ACM Comput. Surv*, 41, 2, Article 10 ,February 2009.

[2] G.Tsatsaronis, I.Varlamis, K.Nørvag," An Experimental Study on Unsupervised Graph-based Word Sense Disambiguation", *CICLing - Conference on Intelligent Text Processing and Computational Linguistics*2010.

[3] M. Lesk." Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone". In *Proc. of the SIGDOC Conference*, pages 24–26,1986.

[4] R.Navigli, P.Velardi. "Structural semantic interconnections: A knowledge-based approach toword sense disambiguation," *IEEE Trans. Pattern Anal. Mach. Intell., 27(7):1075–1086*, 2005.

[5] R.Navigli, M.Lapata, ," An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation," *Pattern Analysis and Machine Intelligence*, IEEE Transactions on, IEEE Computer Society, April 2010.

[6] R.Sinha, R.Mihalcea," Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity," In *IEEE International Conference on Semantic Computing*, ICSC 2007.

[7] R.Navigli, M.Lapata." Graph connectivity measures for unsupervised word sense disambiguation," *IJCAI'07 Proceedings of the 20th international joint conference on Artifical intelligence*,2007.

[8] S.Banerjee, T.Pedersen," An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet ," *CICLing '02 Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*,2002.

[9] M. Palmer, C. Fellbaum, S. Cotton, L. Delfs, and H. Dang. "English tasks: all-words and verb lexical sample". In *Proceedings of ACL/SIGLEX Senseval-2*, Toulouse, France,2001.

[10] B. Snyder and M. Palmer. "The English all-words task". In *Proceedings of ACL/SIGLEX Senseval-3*, Barcelona, Spain,July 2004.

[11] E. Agirre and A. Soroa. "Personalizing pagerank for word sense disambiguation". In *Proc. Of EACL*, pages 33–41, 2009.

[12] Vasilescu, Florentina, P.Langlais, G.Lapalme," Evaluating variants of the Lesk approach for disambiguating words", *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal,633–636. 2004

[13] R. Mihalcea. "Large vocabulary unsupervised word sense disambiguation with graph-based algorithms for sequence data labeling". In *Proceedings of the Human Language Technology / Empirical Methods in Natural Language Processing conference*, Vancouver, 2005.

[14] R. Mihalcea, P. Tarau, and E. Figa. "Pagerank on semantic networks with application to word sense disambiguation". In *Proc. of COLING*, 2004.

[15] R. Navigli. "Online word sense disambiguation with structural semantic interconnections". In *Proc. of EACL*, 2006.

[16] G. Tsatsaronis, M. Vazirgiannis, and I. Androutsopoulos. "Word sense disambiguation with spreading activation networks generated from thesauri". In *Proc. of IJCAI*, pages 1725–1730,2007.

Collaborative Personalized Web Recommender System using Entropy based Similarity Measure

Harita Mehta[1], Shveta Kundra Bhatia[2], Punam Bedi[3] and V. S. Dixit[4]

[1] Computer Science Department, Acharya Narender Dev College, University of Delhi, New Delhi 110019, India

[2] Computer Science Department, Swami Sharaddhanand College, University of Delhi, New Delhi 110036, India

[3] Computer Science Department, University of Delhi, Delhi 110007, India

[4] Computer Science Department, Atma Ram Sanatam Dharam College, University of Delhi, New Delhi 110010, India

Abstract

On the internet, web surfers, in the search of information, always strive for recommendations. The solutions for generating recommendations become more difficult because of exponential increase in information domain day by day. In this paper, we have calculated entropy based similarity between users to achieve solution for scalability problem. Using this concept, we have implemented an online user based collaborative web recommender system. In this model based collaborative system, the user session is divided into two levels. Entropy is calculated at both the levels. It is shown that from the set of valuable recommenders obtained at level I; only those recommenders having lower entropy at level II than entropy at level I, served as trustworthy recommenders. Finally, top N recommendations are generated from such trustworthy recommenders for an online user.

Keywords: *Collaborative Web Recommender System, Trustworthy users, Entropy based Similarity.*

1. Introduction

A web user is usually surrounded by the large quantity of heterogeneous information available on the dynamic web platform. This information overload makes it crucial for the web user to access personalized information. Thus, there is a need for powerful automated web personalization tools for "Web Recommendation" [1] which is primarily aimed at deriving right information at right time. Web recommender systems analyses web logs in order to infer knowledge from the web surfer's sessions and thereby generate effective recommendations for the surfer. It has been observed that, web surfer prefers to visit a page that was visited by another likeminded person in the recent past. User based Collaborative web recommender systems have the same role as that of such human recommenders [8, 18]. In such systems, a user profile is a vector of items and their ratings, continuously appended as the user interacts with the system over a specified period of time. This user profile is compared with the profiles of other users in order to find overlapping interests among users. Thus, it generates recommendations based on inter user similarity. The idea to use the concept of inter user similarity is that if a user has agreed with his neighbors in the past, he will do so in the future also. Trustworthiness is amount of confidence on each other, which exist among such pair of inter similar users. We put forward that, trustworthiness can be derived using entropy. Recommendations generated from such trustworthy users are always preferred over recommendations generated from an unknown user [2].

The current generation of web recommender systems, still require further improvements in order to make recommendation method more effective [6]. One of them is "Scalability problem". In order to find users with similar tastes, these systems require data from a large number of users before being effective, and as well as require a large amount of data from each user. Thus, the computational resources required to find inter similar users become a critical issue. In this paper, we propose a method to select trustworthy recommenders from the list of similar users. It is assumed that similar users are valuable users. We put forward that, trustworthiness between similar users can be

calculated on the basis of entropy existing between them. The step II of the proposed algorithm runs at two levels, thereby selecting only trustworthy recommenders in order to reduce computational resources which would be required at the time of generation of recommendations. Entropy [19] is the measure of inter user similarity that exists during recommendation generation process. It is expressed in terms of discrete set of probabilities as given in Eq. (1).

$$H(D(U_t, U_X)) = -\sum_{i=1}^{n} p(d_i) \log_2 p(d_i) \quad (1)$$

where, $D(U_t, U_x)$ is the difference score rating between the target user U_t and user U_x for n unique URLs and $p(d_i)$ is the probability density function of difference score rating. These probabilities depict the degree to which the target user U_t is similar to user U_x. Lower the entropy, higher the degree of inter user similarity. The paper is organized in the following sections. Section 2, emphasizes on past research on similar work. Section 3, discusses the proposed model for collaborative web recommender system followed by experimental study in Section 4. Section 5, concludes the proposed work.

2. Related Work

In collaborative filtering approaches, the system requires access to the item and user identifiers [5, 11]. A simple approach in this family, commonly referred to as user based collaborative filtering [16], creates a social network of users who share same rating patterns with the target user. This network of users is based on the similarity of observed preferences between these users and the target user. Then, items that were preferred by users in the social network are recommended to the target user. Item based collaborative filtering [13], recommends such items to the target user that were preferred by those other users who preferred the same set of items that were preferred by the target user in the past. In many applications, collaborative recommender systems adapt their behavior to individual users by learning their tastes during the interaction in order to construct a user profile that can later be exploited to select relevant items. User's interest are gathered in an explicit way (such as asking user to rate an item on a scale , rank items as per favorite or choosing one item out of many) or implicit way (such as keeping track of item's user views, keeping the list of items purchased in past , analyze users social network & discover similar likes or dislikes.) It is preferred to work with rating data generated implicitly from user's actions rather than explicit collection [15]. Logs of web browsing or records of product purchases, are as an implicit indication for positive opinions over the items that were visited or purchased. There have been many collaborative systems developed in the academia and in the industry. Some of the most important systems using this technique are group lens / Net perception [17], Ringo / Firefly [6], Tapestry [20], Recommender [8]. Other examples of collaborative recommended system include the book recommender system from amazan.com, the PHOAKS system that helps people find relevant information on WWW [14], and the Jester system that recommends jokes [12]. According to [11], algorithms for collaborative recommendation can be grouped into two general classes: memory based (heuristic based) and model based. Memory based collaborative filtering systems compare users against each other directly using correlation or other similarity measures such as scalar product similarity, cosine similarity and adjusted cosine similarly measure. Model based collaborative filtering systems derive a model from historical rating data and use it to make predictions. In the proposed work, we are concentrating on model based user collaborative filtering system. The researchers are trying to improve the prediction accuracy of generated recommendations. Recommendations generated by trustworthy users are preferred over recommendations generated by unknown web surfers [3, 4]. Trustworthiness is the level of satisfaction which the user gets from another user. This has originated an emergent need to measure inter user similarity with respect to trustworthiness among them. In collaborative web recommender systems, inter user similarity can be measured using information entropy which can reduce prediction error in these systems. Means Absolute Error (MAE) is applied to measure the accuracy of recommendations. Lower MAE values represent higher recommendation accuracy [7, 9, and 10]. In [10], by use of similarity measure using weighted difference entropy, it was shown that the quality of recommendation was improved together with reduced MAE. In [9], another similarity weighting method using information entropy was used and showed reduction in MAE and was found to be robust for sparse dataset. In [7], entropy based collaborative filtering algorithm provided better recommendation quality than user based algorithm and achieved recommendation accuracy comparable to the item based algorithm. In our research, the proposed model measures entropy at two levels of a user session to find trustworthy users and generate recommendations with their degree of importance [3, 4] only from such trustworthy users and serve as a means to reduce scalability problem that hampered traditional collaborative filtering techniques.

3. Collaborative Personalized Web Recommender System using Entropy based Similarity Measure

The architecture of a "Collaborative Personalized Web Recommender System using Entropy based Similarity Measure" is proposed in figure 1. In our study, online recommendations are generated for the demo version of the website available at http://www.vtsns.edu.rs. On the request of online user, top N recommendations are generated by the proposed web recommender system. The main components of this system are Interface Unit, Offline Unit and Online Recommendation Generator.

Online user and the recommender system are two basic entities in any recommendation generation process. Interface unit acts as an interface between these entities. It fetches click stream pattern (pages visited by the user) from the current session of online user. It sends the request to the online recommender generator where top N recommendations are furnished for the online user. Finally, the interface unit accepts the generated recommendations and passes it to the browser, so that these recommendations can be displayed for the online user during his/her current session.

Offline unit is the heart of the proposed recommender system. Creation of the knowledge base for online recommender generator rests on this unit. The processor of the offline unit takes web log of the demo site as input and generates recommendations in offline mode for the user patterns stored in the web log. The processor is the backbone of the offline unit and runs in three steps as discussed below.

3.1 Data Preparation (Step I)

Relevant user sessions in the form of Page View (PV) binary matrix are obtained from the raw web log file with the help of pre processing tools i.e. Sawmill [21]. Binary cell value as "1" in the matrix depicts that the page P_m has been accessed in the session id S_n whereas "0" depicts that the P_m has not been accessed in the session id S_n. The PV matrix is split into Training PV matrix (T_1) and Test PV matrix (T_2); Training PV matrix (T_1) is further split into training level I matrix (M_1) and training level II matrix (M_2) which are required inputs for Step II of the processor.

3.2 Selection of Trustworthy Recommenders based on Entropy between the users (Step II)

This step can be broken down into two levels. At level I, Training level I matrix (M_1) is given as input. After initializing users, difference score is calculated between target user and all other users using Eq. (2).

$$D(U_t,U_x) = \left[\left| PV_{(U_t,P_1)} - PV_{(U_x,P_1)} \right| , \cdots \left| PV_{(U_t,P_n)} - PV_{(U_x,P_n)} \right| \right] \quad (2)$$

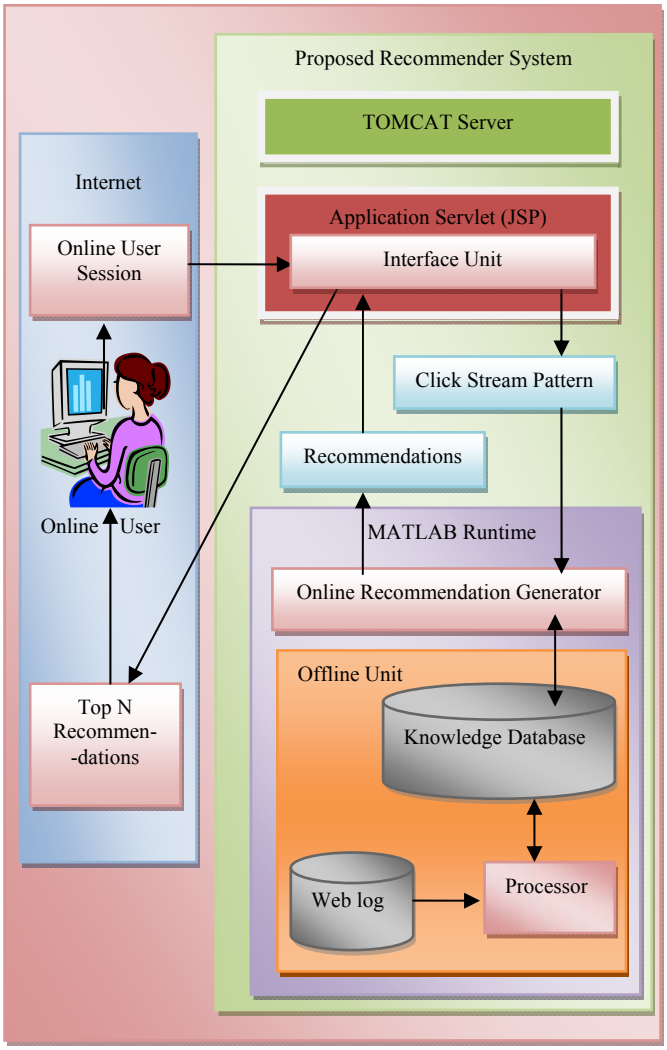


Fig. 1 Architecture of Proposed Recommender System

where, each term represents the page view status of user U_k for page P_n. And, the absolute difference of the page view status for page P_i of two users is considered as (d_i) which is "0" when target user U_t and user U_x have both either viewed or have not viewed the page P_i. Parameter β introduced in Eq. (3) is a similarity threshold which is used to declare whether a web user is a valuable recommender for the target user or not.

$$DZeroCount(U_t,U_x) \geq \beta \times length(D(U_t,U_x)) \quad (3)$$

Here, we count the number of non-zero (d_i) in each (U_t,U_x) pair. If this number is greater than or equal to β times of total number of (d_i) present, then we declare user U_x to be a valuable recommender for target user U_t. Level I entropy (E_1) among such pair of valuable recommenders is

calculated using Eq. (1) and for each target user, list of valuable recommenders arranged in descending order of level I entropy is produced. This set of valuable recommenders for all the users and Training level II matrix (M_2) are given as input at level II. For such valuable users, level II entropy (E_{II}) is calculated using Eq. (1). Lower the entropy, higher the inter-user similarity. If level I entropy is less than level II entropy, then inter user similarity is more. It depicts that the interest of the user remains similar to that of the target user. So the user is considered as trustworthy user for the target user. For such pairs, actual entropy (E_A) is obtained using Eq. (4).

$$E_A(U_t,U_x) = \left(E_I(U_t,U_x) - E_{II}(U_t,U_x)\right)/2 \quad (4)$$

Finally, list of trustworthy recommenders arranged in descending order of actual entropy is produced because lower the entropy, higher the similarity. The algorithm is depicted in figure 2 and figure 3. Our approach reflects a solution to scalability problem in step II, by reducing computational resources; since it generates recommendations only from trustworthy recommenders.

3.3 Generation of Recommended Pages with their degree of importance (Step III)

Set of trustworthy recommenders for all users (R_T) obtained from Step II along with Page View matrix (PV matrix) prepared in Step I and Page visit frequency count (total number of users who have accessed that page) are given as input. Those pages which have not been visited by the target user, but have been visited by its trustworthy recommender, are considered as recommended page for the target user. Finally, evaluate degree of importance for generated recommendations using Eq. (5). Algorithm is depicted in figure 4.

$$DOI(U_t,P_{rec}) = \left(1 - E_c/T_c\right) \times F_c \quad (5)$$

where, T_c is the total number of trustworthy recommenders who have recommended the page P_{rec} to the target user U_t, F_c is the total number of users who have viewed the page P_{rec} and E_c is the total actual entropy value that the trustworthy users have assigned to the page P_{rec}. These recommendations generated by the processor in the offline mode act as knowledge base for the online recommendation generator. The knowledge base constructed by the offline unit consists of user click stream patterns and their recommended pages. Ongoing session information of the online user captured by the interface unit is given as input to the online recommendation generator. It matches the online partial click stream pattern of the online user with the partial click stream patterns of same length stored in knowledge base and finds trustworthy recommenders. Finally, Top N recommendations from the set of these trustworthy recommenders are provided to the interface unit. The algorithm is shown in figure 5.

Input: Training Level I Matrix (M_1).
Process:
Step1:
Do
  Let the first user in M_1 be the target user U_t.
  Let the second user in M_1 be user U_x.
  Evaluate D (U_t, U_x) using Eq. (2).
  Using Eq. (3), if U_x is a valuable user for the target user U_t then
    Append the pair (U_t, U_x) to R_V.
    Calculate $E_I(U_t,U_x)$ using Eq. (1)
Repeat for all possible user pairs in M_1
Step 2:
For each target user
  Arrange the list R_V in descending order of $E_I(U_t,U_x)$
End for
Output: Set of valuable recommenders for all users (R_V).

Fig. 2 Algorithm for Level I

Input: Set of valuable recommenders for all users (R_V).
  Training Level II Matrix (M_2).
Process:
Step1:
Do
  Let the first user in M_2 which is also present in R_V be target user U_t.
  Let second user in M_2 which is also present in R_V be user U_x.
  Evaluate $D(U_t,U_x)$ using Eq. (2)
  Calculate $E_{II}(U_t,U_x)$ using Eq. (1)
Repeat for only those pair of users in M_2 which are also present in R_V.
Step 2:
For all the users in R_V,
  If $E_I(U_t,U_x)$ <= $E_I(U_t,U_x)$, then
    Calculate $E_A(U_t,U_x)$ using Eq. (4)
    Append the pair (U_t,U_x) to R_T.
Step3:
For each target user
  Arrange the list R_T in the descending order of $E_A(U_t,U_x)$.
End for
Output: Set of trustworthy recommenders for all users (R_T).

Fig. 3 Algorithm for Level II

Input: Set of trustworthy recommenders for all users (R_T).
  Page View Matrix (PV Matrix).
  Page Frequency Count Matrix (FC Matrix).
Process:
Step1:
  Do
    Let the target user U_t be a user in R_V
    Let user U_x be the recommended user of U_t in R_V.
    From PV Matrix,
      Find those pages that have been viewed by user U_x and have not been viewed by target user U_t.
      Store such pages in Recommended Page List (Rpage).
    Do
      Evaluate $DOI(U_t,P_{rec})$ using Eq. (5)
    Repeat for each Recommended Page in Rpage.
    Arrange in decreasing order of $DOI(U_t,P_{rec})$
  Repeat for all pair of users in R_T.
Step 2:
Extract Top N Recommendations for each target user U_t where N = {2, 3, 5, 10}
Output: Set of Recommended Pages with their degree of importance for all the users in PV Matrix.

Fig. 4 Algorithm for Offline Unit

Input: Set of Recommended Pages with their degree of importance for all the users in PV Matrix.
Ongoing Session of online target User.

Process:

- Read ongoing user session information from csv files.
- Generate 43 bit binary user click stream pattern where a high at bit position 'P' signifies that the user has visited URL 'P'. (Assume that 43 unique URLs are numbered from 1 to 43)
- Obtain the last visited URL position (say P_j) and split the 43 bit binary user click stream pattern into two parts.($P_1,P_2,\ldots..P_j$) and ($P_{j+1},P_{j+2},\ldots P_{43}$)
- The unique click pattern stored in the knowledge base is split into two partial patterns of same dimensions. (For example,$K_1,K_2,\ldots K_{43}$ is split into $K_1,K_2,\ldots..K_j$ and $K_{j+1},K_{j+2},\ldots K_{43}$).
- Compare online user's click stream pattern $P_1,P_2,\ldots..P_j$ with all the partial patterns $K_1,K_2,\ldots..K_j$ stored in the knowledge base and obtain those trustworthy users who are above the similarity threshold.
- From the set of these trustworthy users, top N Recommendations are fetched and given to the interface unit.

Output: Set of Top N Recommended Pages for online target user.

Fig. 5 Algorithm for Online Recommendation Generator

4. Experimental Study

4.1 Dataset

The demo version of the website (http://www.vtsns.edu.rs) was prepared using Microsoft front page. A prototype of the proposed system was implemented using MATLAB software [22] with TOMCAT server [24] on JAVA platform [23]. The internet platform was realized based on Java Server Pages (JSP) with Tomcat Server as servlet container. Here, Tomcat server acted as a container of the system servlet. The servlet itself was written in JSP and was run on Matlab software. The online target user with the help of web browser got connected with the server via internet provided by Tomcat server. The demo version of the website viewed by the online target user was displayed by the application servlet written in JSP. This servlet gathered the click stream of the online target user via web browser and sent it to the Matlab runtime library. The online recommendations generated were returned by Matlab to the application servlet which displayed them in the demo site via web browser. The experiment proceeded in a desktop PC environment consisting of Intel Core 2 Duo @ 3.00GHz and 2GB RAM. A web usage log file (http://www.vtsns.edu.rs/maja/vtsnsNov16) containing 5999 web requests to an institution's official website on November 16, 2009 was used as dataset. Sawmill processed these requests and grouped the hits into initial sessions based on the visitor id by assuming that each visitor contributes to a session. A session timeout interval of 30 minutes was considered for generating final sessions and sessions longer than 2 hours were eliminated. Page view count is the number of pages accessed by the user. Average page view count obtained from page view matrix was 5.4. So, we optimized our matrix by deleting those sessions that had visited less than 5 pages and deleted those URLs which were visited in only one or two sessions. Finally, we obtained 122 sessions with 43 unique URLs which was used as the input to verify the proposed recommendation generation process. Table 1 shows sample data of 5 users. We considered 80% of the dataset as training page view matrix and rest 20% as test page view matrix. Further, for calculating entropy at two levels, training page view matrix was split vertically with 22 pages at level I and rest 21 pages at level II. We assumed similarity threshold $\beta = 80\%$.

Table 1: Page View Matrix (sample data for 5 users)

Page/ User	***P1***	***P2***	***….***	***….***	***P41***	***P42***	***P43***
U1	1	0	….	….	0	1	1
U2	1	1	….	….	1	0	1
U3	1	1	….	….	0	0	0
U4	1	0	….	….	1	0	1
U5	0	0	….	….	0	1	0

4.2 Results

Figure 6(a) obtained after running step I and II shows a graph depicting number of valuable and trustworthy recommenders. It can be clearly seen that the number of valuable recommenders obtained at level I are considerably decreased at level II to obtain number of trustworthy recommenders. For example, in case of user U_{42}, out of 36 valuable users only 28 users are trustworthy and in case of user U_{43}, the number reduced from 11 to 7.

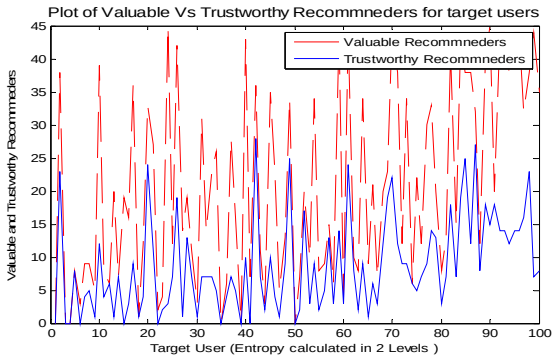


Fig. 6(a) Graph depicting Number of Valuable and Trustworthy Recommenders for 100 users

Figure 6(b) shows the priority of the trustworthy users for user U_{43}. Figure 7 shows plot of level I entropy and level II entropy for trustworthy users of target user U_{43}. It can be clearly visualized that, user U_{97} has higher priority than user U_{88} because the difference between level I entropy and level II entropy is lesser than that for user U_{88}. Finally, recommendations were generated at step III and Figure 8 depicts degree of importance of recommended pages for target user U_{43}.

Target User	Trust Worthy User	Level One Entropy	Level Two Entropy	Actual Entropy
42	40	0.8108	0.2092	0.3008
43	97	0.8108	0.6275	0.0917
43	38	0.6081	0.4183	0.0949
43	52	0.6081	0.4183	0.0949
43	66	0.6081	0.4183	0.0949
43	79	0.6081	0.4183	0.0949
43	88	0.4054	0.2092	0.0981
43	7	0.8108	0.4183	0.1962
44	80	0.8108	0.6275	0.0917
44	58	0.6081	0.4183	0.0949
45	17	0.8108	0.6275	0.0917
45	40	0.8108	0.6275	0.0917

Fig. 6(b) Prioritized trustworthy users for user U_{43}.

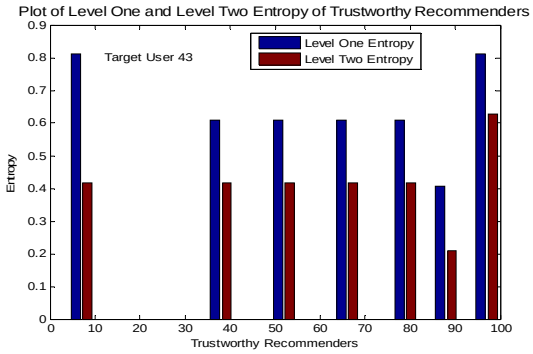


Fig. 7 Plot of Level I Entropy and Level II Entropy of Trustworthy Recommenders for the target user U_{43}.

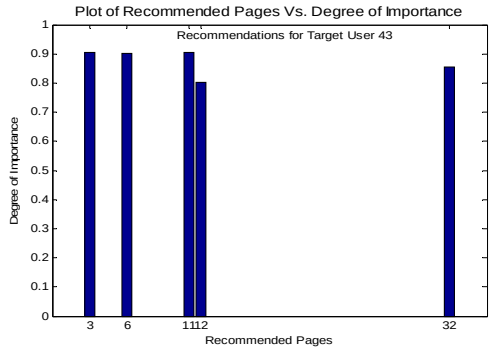


Fig. 8 Plot of Recommended Pages vs. Degree of Importance for target User U_{43}.

We conducted a set of experiments to better understand how entropy calculated at levels improves the selection of trustworthy users. At step III, top N recommendations with their degree of importance was obtained where N= {2, 3, 5, 10}. To check the efficiency in offline mode, it was assumed that recommendations were generated after the target user has visited at least 6 URLs on the website. In order to find out similar users for the target user, similarity threshold β was set to 50% (i.e. at least 3 similar clicks). For this purpose, training page view matrix was split into visited training PV matrix (containing those 6 URLs already visited) and unvisited training PV matrix (rest of unvisited URLs). Similarly, test page view matrix was split into visited test PV matrix and unvisited test PV matrix of same dimensions. For each target user in visited PV test matrix, similar users were identified from visited PV training matrix. From the list of top N recommendations generated at step III, recommendations were obtained for these similar users and were stored in the predicted list. Finally, from the unvisited test PV matrix, actual pages viewed were found and stored in the actual list. In this experiment, we used Means Absolute Error (MAE), a statistical accuracy metric. Suppose, the set of entropy values predicted from the training set is $\{p_1, p_2,\ldots p_n\}$, and the corresponding set of actual entropy values from the test set is $\{q_1,q_2,\ldots q_n\}$then MAE is obtained using Eq. (6).

$$MAE = \frac{\sum_{i=1}^{n}\left(|p_i - q_i|\right)}{N} \qquad (6)$$

where, p_i is the predicted entropy, q_i is the actual entropy, n is total number of URLs and N is total number of levels . Lower MAE values represent higher trust value between the pair of users because the interests of these users remain same throughout the session. MAE values obtained for Top N recommendation sizes are shown in table 2. The graph in figure 9 depicts that for all recommendation sizes, MAE values remained less than 0.5.

Table 2: MAE for Top N Recommendations

Top N	***Top 2***	***Top 3***	***Top 5***	***Top 10***
Proposed System	0.2114	0.2591	0.3245	0.4027

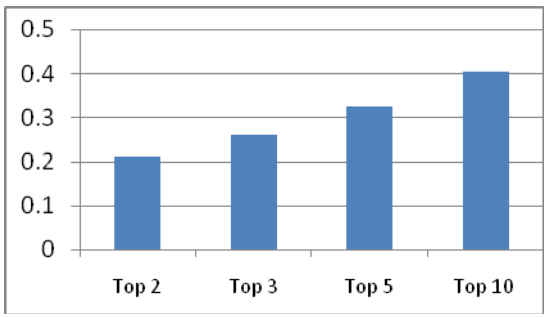


Fig 9 MAE for Top N Recommendations

To measure the quality of the proposed recommender system, two information retrieval measures, Precision and Recall were studied. Precision is the proportion of recommendations that are good recommendations and recall is the proportion of good recommendations that appear in top recommendations. Suppose, the set of URLs that are viewed by the target user are Relevant URLs and those URLs that are recommended by the recommender are Retrieved URLs, then precision ratio and recall ratio are obtained using Eq. (7a) and (7b) respectively.

$$\text{Precision} = \frac{|(\text{Relevant URLs}) \cap (\text{Retreived URLs})|}{|(\text{Retreived URLs})|} \quad (7a)$$

$$\text{Recall} = \frac{|(\text{Relevant URLs}) \cap (\text{Retreived URLs})|}{|(\text{Relevant URLs})|} \quad (7b)$$

One cannot achieve 100% Precision ratio or Recall ratio. So, we understand them relatively in relation to other systems. For this comparison, we prepared a single level entropy based recommender system (In introduction, we argued that algorithm will run in two levels at step II in order to generate recommendations only from trustworthy recommenders thereby reducing the required computational resources. To prove the statement, we implemented another Single Level Entropy based algorithm (SLE Web Recommender), in which the dataset was not divided into two sessions. It selected valuable users for a target user implicitly based on inter user difference score similarity obtained from page view matrix. Further, entropy for such valuable recommenders was calculated from the entire dataset. Similarity threshold was set to half of difference of maximum entropy and minimum entropy of the system. Those valuable recommenders who had entropy less than similarity threshold were considered as trustworthy recommenders. Finally, from such trustworthy recommenders, recommended pages with their degree of importance were obtained.) SLE web recommender was compared with our proposed web recommender system. Precision and Recall ratios recorded at various recommendation sizes is shown in table 3(a) and 3(b) respectively. Further, corresponding graphs are shown in figure 10(a) and 10(b). From this viewpoint, the measurements of our system showed better performance in both precision and recall ratios. It can be clearly seen that as the recommendation size increases, precision ratio decreases whereas recall ratio increases. For top 5 recommendation size, precision ratio marginally increased but recall increased almost 2.2 folds (i.e. from 24.1 % to 53.1%). Also, for top 10 recommendation size, recall increased almost 2.5 folds (i.e. from 24.1 % to 61.9%) and precision ratio marginally increased. Recall measures may be improved by increasing the recommendation size; however, it is best not to recommend too many items to users in order to avoid overloading. Choosing a proper recommendation size will be an appropriate topic for future studies.

Table 3(a): Precision Ratios

Top N	*Top 2*	*Top 3*	*Top 5*	*Top 10*
SLE Web Recommender	19.90%	18.10%	18.10%	18.00%
Proposed Web Recommender	30.30%	27.10%	24.50%	22.10%

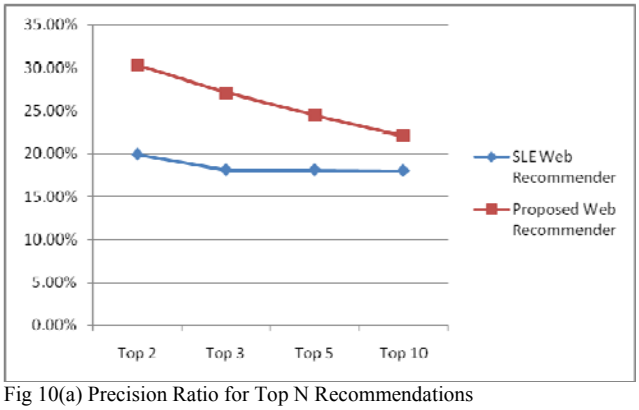


Fig 10(a) Precision Ratio for Top N Recommendations

Table 3(b): Recall Ratios

Top N	*Top 2*	*Top 3*	*Top 5*	*Top 10*
SLE Web Recommender	21.50%	23.70%	24.10%	24.10%
Proposed Web Recommender	30.10%	38.70%	53.10%	61.90%

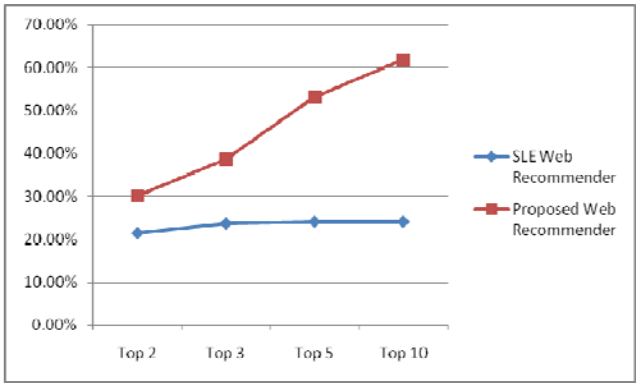


Fig 10(b) Recall Ratio for Top N Recommendations

After running steps I to III in offline mode, knowledge database was created. The database contained unique click patterns and their recommended pages. A demo version of the site available at http://www.vtsns.edu.rs was developed. Figure 11(a) shows the snapshot of the demo site. The snapshot of top N recommendations generated

online is displayed in figure 11(b). Top N Recommendations were generated for an online user and similarity threshold β was set to 50%.

Fig. 11(a) Snap Shot of Demo Site

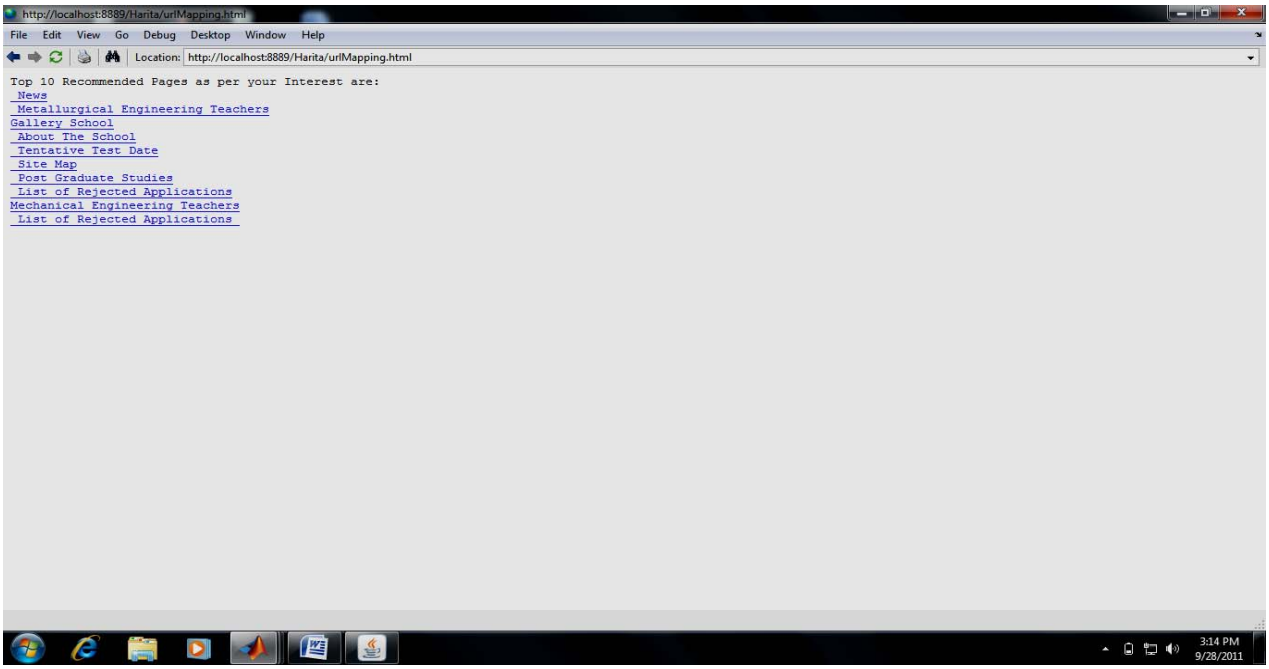


Fig. 11(b) Snap Shot of Top 10 Recommendations for an online user

5. Conclusions

The interest in the area of collaborative web recommender system still remains high because of the abundance of practical applications that demands personalized recommendations. In this paper, a "Collaborative Personalized Web Recommender System using Entropy based Similarity" is implemented in order to solve the problem of scalability. Traditionally, collaborative systems have relied heavily on inter user similarity based on difference score rating. We have argued that the difference score similarity on its own may not be sufficient to generate effective recommendations. Specifically, we have introduced the notion of entropy in reference to degree to which one might trust a specific user during recommendation generation. We have developed entropy based computational model which operated at two levels instead of single level. At both levels, recommenders were generated by monitoring entropy between similar users based on difference score rating. We have described a way to suppress the generation of recommenders who were valuable but not trustworthy. We found that the use of entropy at two levels had a positive impact in solving scalability problem. Top N recommendations were generated and MAE was found to be less than 0.5 for all recommendation sizes. As the recommendation size increased, Precision ratio decreased and recall ratio increased. Precision and recall for top N recommendations were found to be better when compared with single level entropy based web recommender system.

References

[1] G. Adomavicius, and A. Tuzhilin, "Toward the next generation of recommender system; a survey of the state of the art and possible extensions", IEEE transaction on knowledge and data engineering, June 2005, Vol 17, No. 6.

[2] P. Bedi, and H. Kaur, "Trust based personalized recommender system", INFOCOMP Journal of Computer Science. 5(1), 2006 pp. 19-26.

[3] B. Gupta, P. Bedi and H. Kaur Negi, "Trust based personalized ecommerce system for farmers", In proceedings IICAI'09-4th Indian International Conference on Artificial Intelligence, December 16-18, 2009, Tumkur, pp. 1085-1093.

[4] P. Bedi, and S. Aggarwal, "Influence of terrain on modern combat: Trust based recommender system", Defence Science Journal, Vol 60, no. 4, July 2010, pp. 405-411.

[5] D. Billsus, and M. Pazzani, " Learning collaborative information filters", Proceedings of Fifteenth International Conference on Machine Learning, Madison, Wisconson, USA, 1998, pp. 46-54.

[6] R. Bruke, "Hybrid recommended systems; Survey and experiments", user modeling and user adapted interaction, 2002, Vol 12, pp. 331-370.

[7] C. H. Piao, J. Zhao, and L. J. Zheng, "Research on entropy based collaborative filtering algorithm and personalized recommendation in e- commerce", SOCA, 2009, Vol 3, pp. 147-157, Springer.

[8] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using calloborative fittering to weave an information tapestry", Comm. ACM, 1992, Vol 35 number 12 , pp. 61-70.
[9] H. J. Kwon, T. H. Lee, J. H. Kim and K. S. Hong, "Improving Prediction accuracy using entropy weighting in collaborative filtering", Symposia and workshops on Ubiquitous, Automatic and Trusted Computing, 2009 IEEE.
[10] H. J. Kwon, T. H. Lee and K. S. Hong, "Improving memory based collaborative filtering using entropy based similarity measures", In proceedings of International Symposium on web information systems and applications (WISA '09), 2009 Academy Publisher.
[11] J. S. Breese, D. Heckerman, and C. M. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering", In UAI: Uncertainty in Artificial Intelligence, 1998, pp. 43–52.
[12] K. Goldberg, T. Roeder, D. Gupta and C. Perkins, "Eigentaste: a constant time collaborative filtering algorithm", information retrieval , july 2001, Vol 4, No. 2, pp. 13 3-151.
[13] Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering", IEEE Internet Computing, 2003, Vol 7 (1).
[14] L. Terveen, W. Hill, B. Amento, D.Mc Donald and J. Creter, "PHOAKS: a system for sharing recommendation", Comm. ACM, 1997, Vol 40, No 3, pp. 59-62.
[15] M. Claypool, P. Le, M. Waseda, and D. Brown, "Implicit interest indicators", In Intelligent User Interfaces, ACM Press, 2001, pp. 33–40.
[16] M. McNee, J. Riedl, and J. K. Konstan, "Making recommendations better: an analytic model for human-recommender interaction", In CHI '06 Extended Abstracts on Human Factors in Computing Systems, 2006.
[17] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "Group lens; an open architecture for collaborative filtering of networks", in proceedings of the conference on computer supported co-operative work, chapel hill, NC,1994, pp. 175-186.
[18] P. Resnick, and H. R. Varian , "Recommender System", communication of the ACM, 1997, Vol. 40 No. 3, pp. 56-58.
[19] Shannon, and E. Claude, "Prediction and Entropy of printed English", The Bell System Technical Journal, Vol 30, 1951, pp 50-64.
[20] U. Shardanand, and P. Maes, " Social information filtering; algorithm for automating word of mouth" in CHI'95 conference proceeding on human factor in computing system, denver, 1995, pp. 210 - 217.
[21] Sawmill http://sawmill.net
[22] Matlab http://www.mathworks.com
[23] Java http://java.sun.com
[24] Tomcat http://tomcat.apache.org

Ms. Harita Mehta is a Research Scholar and working as an Assistant Professor in the Department of Computer Science, Acharya Narender Dev College, University of Delhi. Her research area is Web Recommender Systems and is currently pursuing PhD under Dr. V.S. Dixit from Department of Computer Science, University of Delhi.

Ms. Shveta Kundra Bhatia is a Research Scholar and working as an Assistant Professor in the Department Of Computer Science, Swami Sharaddhanand College, University of Delhi. Her research area is Web Usage Mining and is currently pursuing PhD under Dr. V.S. Dixit from Department of Computer Science, University of Delhi.

Dr. Punam Bedi received her Ph.D. in Computer Science from the Department of Computer Science, University of Delhi, India in 1999 and her M.Tech. in Computer Science from IIT Delhi, India in 1986. She is an Associate Professor in the Department of Computer Science, University of Delhi. She has about 25 years of teaching and research experience and has published more than 110 research papers in National/International Journals/Conferences. Dr. Bedi is a member of AAAI, ACM, senior member of IEEE, and life member of Computer Society of India. Her research interests include Web Intelligence, Soft Computing, Semantic Web, Multi-agent Systems, Intelligent Information Systems, Intelligent Software Engineering, Software Security, Intelligent User Interfaces, Requirement Engineering, Human Computer Interaction (HCI), Trust, Information Retrieval, Personalization, Steganography and Steganalysis.

Dr. V. S. Dixit is working as senior Assistant Professor in the Department Of Computer Science, AtmaRam Sanatam Dharam College, University of Delhi. His research area is Queuing theory, Peer to Peer systems, Web Usage Mining and Web Recommender systems. He is currently engaged in supervising the research scholars (Ms Harita Mehta and Ms Shveta Kundra Bhatia) for PhD. He is Life member of IETE.

Hierarchal Object Oriented Fault Tolerant Secured and Atomic Mobile Agent Model.

Mayank Aggarwal[1] and Nipur[2]

[1] Department of Computer Science and Engineering, Gurukul Kangri University, Haridwar, Uttarakhand 249404, India

[2] Department of Computer Science, Kanya Gurukul Mahavidyalaya, Dehradun, Uttarakhand ZIP/Zone, India

Abstract

Mobile Agents are soft wares migrating from one node to another to fulfill the task of its owner. Mobility introduces two major challenges in front of mobile agent namely reliability and security. As the agent moves from one node to another the goal to complete its task safely is difficult to achieve. Mobile agents are no longer a theoretical concept, much architecture for their realizations have been proposed. However, it has to be confirmed that any failures (machine or agent) do not lead to blocking of agent together with the security issues. This paper proposes a model which deals with both the problems; Fault Tolerance and Security, further it also adds atomicity to mobile agents execution i.e. either all the goals are achieved or none is achieved. This paper proposes a Hierarchal model which uses the concepts of object oriented technology, grouping, atomicity and authentication to deal with the blocking problem and security issues.

Keywords: Mobile Agent, Blocking, Atomicity, Object Oriented, Grouping.

1. Introduction

Mobile agents are software that acts autonomously on behalf of a user and migrate through a network of heterogeneous machines [1].The advantage for using mobile agent technology is that interaction cost for the agent-owner is remarkably reduced since after leaving its owner the agent migrates from one host to the next autonomously [9]. Still, even today, only few real applications rely on mobile agent technology might be due to the lack of transaction support for mobile agents [6]. When a mobile agent migrates from one host to another variety of faults may occur, it may be a system crash, corruption of agent, failure of platform, link failure etc but the objective should be that execution is not blocked[7]. To evade from blocking problem replication was introduced which provided another challenge of exactly once problem which was tackled in [15]. The paper provides a solution for blocking problem by grouping mobile agent platform which provide same type of services. Security remains the major hurdle in the field of mobile agent, as the agent has to be executed on hosts other than its owner chances of the host being malicious is very prominent. A lot of research issues in the security of mobile agent are discussed in [4, 5]. The model proposed a hierarchal structure with authorization process involved at every step to make the system secure together with a trusted hardware approach for final execution. Atomicity of an agent means that if the owner wants more than one task to be done, and all the tasks are interrelated than the transactions should be committed if and only if all the tasks have been successfully carried out. For example an agent whose task is to buy an airline ticket, book a hotel room, and rent a car at the flight destination. The agent owner, i.e., the person or application that has created the agent, naturally wants all three operations to succeed or none at all. Clearly, the rental car at the destination is of no use if no flight to the destination is available. On the other hand, the airline ticket may be useless if no rental car is available. The mobile agent's operations thus need to execute atomically. The proposed model incorporates atomicity by final commitment to be done separately by trust server when it receives results from all the groups

2. Background

Security of mobile agents involves two main issues, protecting agent against platform and protecting platform from agent. A lot of research has been done to make the mobile agent system secure. Techniques like Software-based Fault Isolation, Safe Code Interpretation, Signed Code, State Appraisal, Path Histories, and Proof Carrying Code has been used to protect the platform [18]. Similarly, various security mechanisms for protecting agent themselves are Partial Result Encapsulation, Mutual Itinerary Recording, Itinerary Recording with Replication and Voting, Execution Tracing, Environmental Key

Generation, Computing with Encrypted Functions, and Obfuscated Code [16, 17] .This proposed model ensures security by several measures like mutual authentication between agent and host, authentication at each level, trusted hardware and the concept of path history [2]. The IEEE83 defines fault tolerance as "The ability of a system or component to continue normal operation despite the presence of hardware or software faults." There are several fault tolerance approaches [7] like Spatial Replication, Primary Backup Protocols, Active Clients Primary Backup Model these approaches result in violation of exactly once property. Some other approaches which preserve exactly once property are Agent execution model, Enhanced Agent execution model, Voting Protocol, Rear Guards etc. The proposed model makes the system fault tolerant (non blocking) by making groups of similar mobile agent platform. The concept of Hierarchal model has been discussed in [10, 12, and 14]. Blocking occurs [7], if the failure of a single component prevents the agent from continuing its execution. In contrast, the non-blocking property ensures that the mobile agent execution can make progress any time. A non-blocking transactional mobile agent execution has the important advantage, that it can make progress despite failures. In a blocking agent execution, progress is only possible when the failed component has recovered. The proposed model incorporates non blocking property by grouping of mobile agent platforms and ensures atomicity by doing the final commitment at the trust server. As far as we have surveyed this is the first model which incorporates security, fault tolerance and also atomicity for transactional mobile agent.

3. Model

Security of mobile agents involve

The proposed model has four main components:

- Trust server
- Local Network Server
- Mobile Agent System (Group)
- Mobile Agent (Object)

2.1 Trust server

Trust server is the topmost layer of the model. It is responsible for mobile agent authentication and commitment. It receives the agent from the group incharge to be migrated to some other group. On receiving the agent it first decrypts the header, which contains the agent id, source id, destination id, path history (if any). The decryption is done by the private key of the server itself. It then checks for any threat in the decrypted header by comparing the information with its knowledge base if the agent is not safe it is put in the prison. After proper authentication, trust server prepares the agent to migrate further. First of all it saves the computed result (if any) in its knowledge base, and then it encrypts the agent with the public key of local network server and sends it to local server.

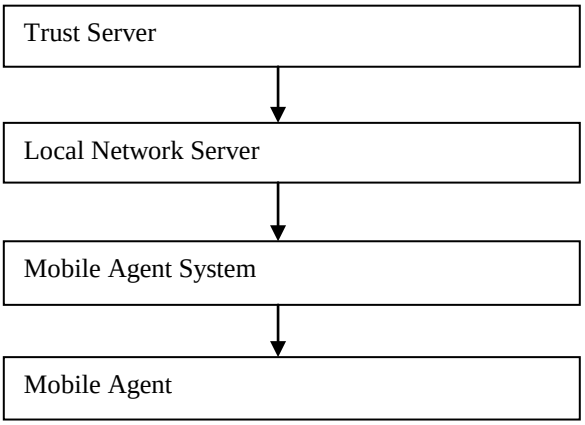


Fig. 1 Hierarchal Model

Further to it, it is also responsible to achieve atomicity, when it has received all the result; it commits all the computed transaction in a safe place protected by the trusted hardware.

2.2 Local Network Server

It receives the agent from the trust server and transfers it to the respective group incharge. On receiving the agent it decrypts the agent, compares the information with its security base and if all is found to be authentic then migrate the agent to the respective group incharge whose id is given in the header. Before sending the agent, it encrypts the agent with the public key of the respective group incharge.

2.3 Group Incharge

The group incharge on receiving the agent decrypts the agent with its private key. Check for the authenticity of the agent and also do the mutual authentication. The group is made of more than one mobile agent platform doing the same type of service. The incharge depending on the load on its members transfer the agent to one of its member for execution. On finishing the task agent returns back o group incharge which encrypt it with public key of trust server

and send it to trust server for further execution. It is designed as System Net.

2.4 Mobile Agent

Mobile agent itself is defined as on object of the Mobile agent platform, which is defined as Object Net of the System Net. The agent is defined as an object having an interface to communicate with outside world, knowledge base, header, path history.

4. Grouping

Blocking is one of the major problems in mobile agent. Many solutions have been presented before to avoid blocking of agents [15]. This paper has done grouping of mobile agent platforms to avoid blocking. A mobile agent submitted at one host within the group can be executed by any host of the group. Grouping can be done based on different criteria like

- Services offered
- Capability of hosts

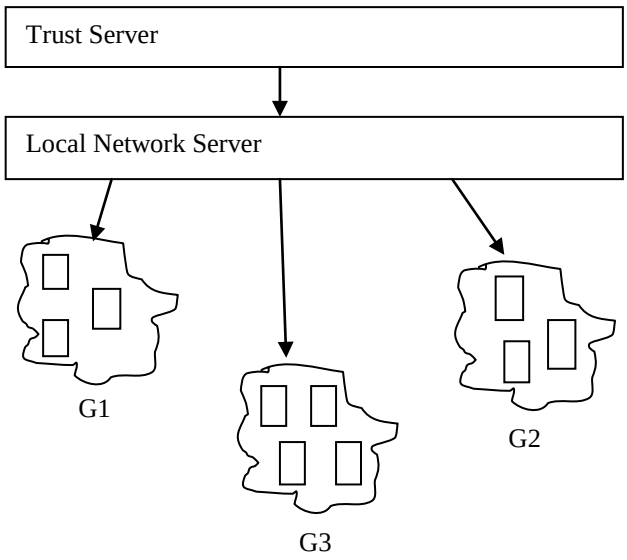


Fig. 2 Detailed Model Showing Groups

- Robustness of the hosts
- Authority to access shared data
- Communication route available between hosts.

This paper has grouped hosts on the basis of the services offered. One of the hosts is decided as incharge of the group. The incharge decide to whom the agent should be sent next within the group. If a particular host fails, agent continues its execution to some other host within the group as decided by the incharge. The problem of replication is avoided as agent is executed on only one host at a time. There may be a case in which incharge itself fails, in that case some other host is nominated as incharge by all other hosts in the group.

5. Hierarchical and Object Oriented Approach

In order to make the model secure and reliable, hierarchal and object oriented approach is used in designing of the model. There are four basic levels in the model, on the top trust sever, below it local server, then mobile agent platform in the form of groups and finally mobile agent which is defined as an object of the mobile agent platform. The hierarchical approach reduces network traffic as well as communication delay. Security is achieved as encryption-decryption is done at each level. Migration of agent from one level to another requires encryption of the agent by the public key of the receiver, which can be decrypted only by the receiver by its private key. Defining agent as an object adds to the security of the agent, it is modified only at the authenticated mobile agent platform. Object oriented approach also supports mobility of the agent.

6. Algorithm

The model has mainly four algorithms one for each level.

6.1 Trust Server Algorithm:

Step 1: Receive the agent

Step 2: Decrypt the agent

Step 3: Authenticate the agent if authenticated go to step 4 else put the agent in prison and exit.

Step 4: Collect the partial results in its knowledge base.

Step 5: If all the results have been collected go to step 6 else go to step 7.

Step 6: Commit the transactions in the assigned safe host.

Step 7: Encrypt the agent with public key of local server.
Step 7: Transfer the agent to local server.

6.2 Local Server Algorithm:

Step 1: Receive the agent

Step 2: Decrypt the agent with private key of local server.

Step 3: Authenticate the agent if authenticated go to step 4 else put the agent in prison and exit.

Step 4: Encrypt the agent with public key of group incharge.

Step 5: Transfer the agent to group incharge.

6.3 Group Incharge Algorithm:

Step 1: Receive the agent

Step 2: Decrypt the agent with its private key.

Step 3: Authenticate the agent if authenticated go to step 4 else put the agent in prison and exit.

Step 4: Search for member host with minimum load

Step 5: Send agent to host for execution

Step 6: If the selected host fails, search for another host and send agent to next selected host.

Step 7: Receive the agent back after it completes its execution.

Step 8: Update the path history, source id, and destination id.

Step 9: Encrypt the agent with public key of trust server.

Step 10: Send the agent to global network.

6.4 Owner Algoritm (Agent):

Step 1: Create the agent

Step 2: Crate header having the source id, destination id, agent id.

Step 3: Encrypt the agent with the public key of Trust Server.

Step 4: Transfer it to trust server.

7. Workflow of Model

The proposed model goes through a sequence of steps to achieve its goal. The model works as described in the above algorithms. The work flow is shown in Figure 3, with numbers showing the sequence of flow of agent.

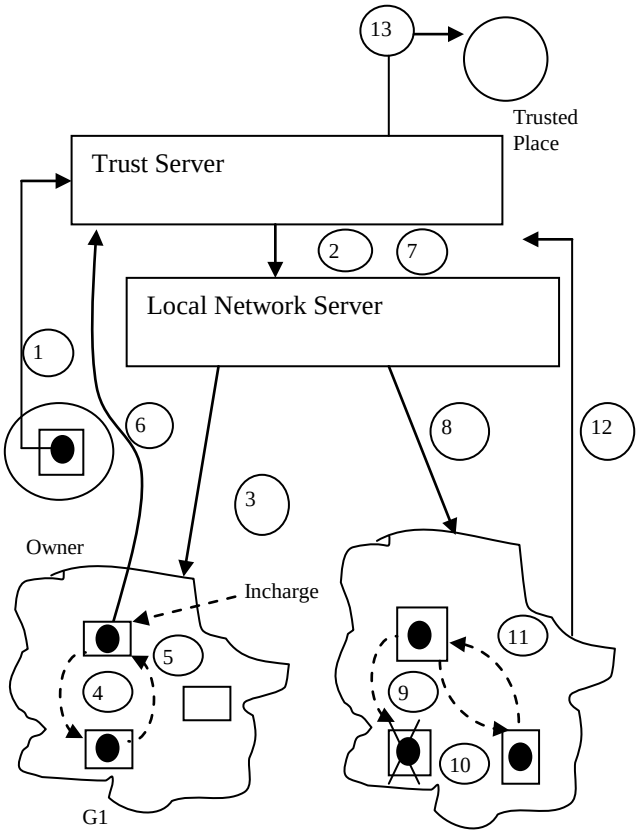


Fig. 3 Workflow of the model

8. Conclusion

The proposed model gives a unique way to implement a secured mobile agent model which tackles the problems of fault also. The simulation of the model is to be done, which is also under process by us using CPN tools [3].

References

[1] Antonio Corradi, Marco Cremonini, Rebecca Montanari, and Cesare Stefanelli, "Mobile agents integrity for electronic commerce applications," Information Systems, 24(6), 1999.

[2] Cao, Chun and Lu Jian, "Path-history-based access control for mobile agents," International Journal of Parallel, Emergent and Distributed Systems, vol 21: 3, pp 215 — 225 , 2006.

[3] CPN Tools website:www.daimi.au.dk/CPNtools

[4] Dirk Westhoff, Markus Schneider, Claus Unger, and Firoz Kaderali, "Methods for protecting a mobile agent's route," Information Security, Second International Workshop (ISW'99), number 1729 in LNCS. Springer Verlag, 1999.
[5] Dirk Westhoff, Markus Schneider, Claus Unger, and Firoz Kaderali, "Protecting a mobile agent's route against collusions," Selected Areas in Cryptography, 6th Annual International Workshop (SAC'99), number 1758 in LNCS. Springer Verlag, 2000.
[6] Dong Chun Lee and Jeom Goo Kim, "Adaptive migration strategy for mobile agents on internet," Technologies for E-Services (TES 2001), Second International Workshop, Proceedings, number 2193 in LNCS. Springer Verlag, 2001.
[7] Fred B. Schneider "Towards fault-tolerant and secure agentry," Distributed Algorithms, 11th International Workshop (WDAG'97), Proceedings, number 1320 in LNCS. Springer Verlag, 1997.
[8] Garrigues, C., et al, "Promoting the development of secure mobile agent applications," J. Syst. Software (2009), doi:10.1016/ j.jss.2009.11.001
[9] Giovanni Vigna, "Cryptographic traces for mobile agents," G. Vigna, editor, Mobile Agents and Security, number 1419 in LNCS. Springer Verlag, 1998.
[10] H.Pathank, K.Garg, "CPN Model for Hierarchical fault tolerance protool for mobile agent systems," 16th IEEE International conference on netwroks, 2008.
[11] Lotfi Benachenhou, Samuel Pierre, "Protection of a mobile agent with a reference clone," Elsevier , Computer Communications , vol 29, pp. 268-278, 2006.
[12] N.Desai, K.Garg, M.Mishra, "Modelling Hierrarchical Mobile Agent Security Prorotocol Using CP Nets," Springer-Verlag Berlin Heidelberg, LNCS 4873, pp 637-649, 2007.
[13] Price, Sean M., "Host-Based Security Challenges and Controls: A Survey of Contemporary Research," Information Security Journal: A Global Perspective, vol 17: 4, pp 170 — 178, 2008.
[14] Satoh I, "A hierarichal model of mobile agents and its multimedia applications," Parallel and distributed systems : Workshops, Seventh International conference , Japan, 2000.
[15] Stefan Pleisch, Andre Schiper, "Modeling Fault-Tolerant Mobile Agent Execution as a Sequence of Agreement Problems," Proceedings of the 19th IEEE Symposium on Reliable Distributed System (SRDS) p.p. 11-20, 2000.
[16] Tomas Sander and Christian F. Tschudin, "Protecting mobile agents against malicious hosts," in G. Vigna, editor, Mobile Agents and Security, number 1419 in LNCS. Springer Verlag, 1998.
[17] Venkatesan S, et al., "Advanced mobile agent security models for code integrity and malicious availability check.," J Network Comput Appl, doi:10.1016/j.jnca.2010.03.010 ,2010 .
[18] W.A.Jansen, "Countermeasures for mobile agent security," Elsevier, Computer Communications, vol. 23 , pp. 1667-1676 , 2000.

Increasing DGPS Navigation Accuracy using Kalman Filter Tuned by Genetic Algorithm

M. R. Mosavi[1], M. Sadeghian[2] and S. Saeidi[3]

**[1] Department of Electrical Engineering, Iran University of Science and Technology
Tehran, 16846-13114, Iran**

**[2] Electronic Education Center, Iran University of Science and Technology
Tehran, 16846-13114, Iran**

**[3] Department of Electrical Engineering, Iran University of Science and Technology
Tehran, 16846-13114, Iran**

Abstract

Global Positioning System (GPS) is being used in aviation, nautical navigation and the orientation ashore. Further, it is used in land surveying and other applications where the determination of the exact position is required. Although GPS is known as a precise positioning system, there are several error sources which are categorized into three main groups including errors related to satellites, propagation and GPS receivers. Regarding wide applications of GPS systems and the importance of its accuracy, these exiting errors could be averted by Differential GPS (DGPS) method. In this paper, a Kalman Filter (KF)-based algorithm which is adapted with Genetic Algorithm (GA) is proposed to reduce errors in GPS receivers. The model's validity is verified by experimental data from an actual data collection. Using the practical implementations the experimental results are provided to illustrate the effectiveness of the model. The experimental results suggest that it is possible to reduce position RMS errors in single-frequency GPS receivers to less than 1 meter. Accordingly, effective error value improves to 0.4873 meter utilizing KF adapted with GA as compared to traditional KF.

Keywords: *Improvement in Accuracy, Differential GPS, Kalman Filter, Genetic Algorithm*

1. Introduction

GPS (Global Positioning System) is a satellite-based positioning and navigating system which is able to determine the instant position, velocity and the time of user on the earth anytime and anywhere. A constellation of at least 24 well-spaced satellites that orbit the Earth makes it possible for people with ground receivers to pinpoint their geographic location. GPS was funded by and controlled by the U. S. Department of Defense (DOD). While there are many thousands of civil users of GPS world-wide, the system was designed and was operated by the U. S. military. Consumer receivers are the approximate size of a hand-held calculator, cost a few hundred dollars, and provide a position accurate to 25 [m] or so. Military versions decode the signal to provide position readings that are more accurate. The exact accuracy is obtained by the military which can be considered as a military secret. GPS satellites are gradually revolutionizing driving, flying, hiking, exploring, rescuing and map making [1].

GPS provides coded satellite signals that can be processed in a GPS receiver, enabling the receiver to compute position (longitude, latitude and altitude), velocity and time. Four GPS satellite signals are used to compute positions in three dimensions and the time offset in the receiver clock. However, GPS receiver as any measurement tool is affected by different kinds of error sources including hardware, environment or atmosphere which can reduce its measurement accuracy [2].

In order to remove positioning errors and achieve more accuracy, DGPS method can be functional. The underlying premise of Differential GPS (DGPS) is that any two receivers that are relatively close to each other will experience similar atmospheric errors. DGPS requires that a GPS receiver be set up on a precisely known location. This GPS receiver is the base or reference station. The base station receiver calculates its position based on satellite signals and compares this location to the known location. The difference is applied to the GPS data recorded by the second GPS receiver, which is known as the roving receiver. The corrected information can be applied to data from the roving receiver in real time in the field using radio signals or through post processing after data capture by special processing software [3]. The problem with this method is slow updating process of differential corrections [4,5].

The purpose of this paper is to represent an algorithm based on Kalman Filter (KF) which is adapted with Genetic Algorithm (GA) in order to reduce errors in GPS receivers. The advantage of the proposed method to traditional KF is its high accuracy. This paper is organized as follows; the sources of errors in GPS systems, DGPS

method, and the process of design and implementation of KF-based estimator are first explained. Then, the optimization method which is used to optimize the KF parameters based on GA is discussed. Next, adopted data collection method and experimental test results, carried out on collected actual data, are reported. Finally, the conclusion is given in the last section.

2. Sources of Errors in GPS

Generally, there can be various factors that affect the quality of the GPS signal and cause calculation errors. These are [6]:

2.1 Errors related to satellites

There are still some errors related to satellites including errors in satellite clock, satellite geometry and also satellite orbits.

Satellite clock: There are two rubidium and two cesium clocks in each satellite responsible for generating GPS satellite signals. These clocks are corrected and adjusted every day by GPS control segment. It should be mentioned that the errors caused by these clocks, working for 24 hours, is equal to 17 [nsec] or 5 [m] in length.

Satellite geometry: In fact, satellite geometry describes the position of the satellites to each other from the view of the ground receiver. For the signals to work properly the satellites have to be placed at wide angles from each other. Poor geometry resulting from tight grouping can result in signal interference [7].

Satellite orbits: Although the satellites are positioned in predetermined orbits precisely, there could be slight changes in their exact positions due to natural phenomena happening in the space. Information about satellites including the orbit data are controlled and updated regularly and are sent to the receivers along with the satellite signals in the package of ephemeris data. Therefore, there can be some differences between the actual position of the satellites and the predicted position.

2.2 Propagation errors

Satellite signals propagate through atmospheric layers during their travel between satellite and GPS receiver. Two significant layers including ionosphere and troposphere cause the satellite signals to slow down as they passes through the atmosphere. However, the GPS system has a built-in model that accounts for an average amount of these disturbances.

Ionosphere layer errors: In the ionosphere developing at a height of 50 to 1000 [km] above the earth, a large number of electrons and positive charged ions are formed by the ionizing force of the sun. There are four conductive layers in the ionosphere which refract the electromagnetic waves from the satellites, resulting in an elongated runtime of the signals. In other words, the delayed coded information of GPS causes pseudo-ranges to appear longer than the real distance from satellite. Ionosphere delays are in connection to frequency and they are mostly corrected by the receiver by calculations. Delays may vary from 5 [m] (at night) to 30 [m] (in day) for low-height satellites and 3-5 [m] for high-height satellites [8].

Troposphere layer errors: The troposphere is the lower part of the earth's atmosphere that encompasses our weather. Tropospheric effect is a further factor elongating the runtime of electromagnetic waves by refraction. The reasons for the refraction are different concentrations of water vapor in the troposphere, caused by different weather conditions. The error caused in this way is smaller than the ionospheric error, but can not be eliminated by calculation. It can only be approximated by a general calculation model.

Multi-path effect: Multi-path is a phenomenon in which a signal of a GPS hits and is reflected off by the surrounding objects like tall buildings, rocks, etc. before being detected by antennas. This causes the signal to be delayed before it reaches the receiver [9].

2.3 GPS receiver errors

There are some errors that originate from measurement processes used in GPS receivers. These errors are referred to as GPS receiver errors. They are mainly connected to the design of antenna, the method of changing analog to digital, band width and calculation algorithms.
Selective Availability (SA) is the intentional degradation of the SPS signals by a time varying bias. SA is controlled by the DOD to limit accuracy for non-U.S military and government users. The potential accuracy of the C/A code of around 30 [m] is reduced to 100 [m] (two standard deviations). For civil GPS receivers, the position determination is less accurate (fluctuation of about 50 [m] during a few minutes). Meanwhile, SA has been permanently deactivated since May 2000 due to the broad distribution and worldwide use of the GPS system [10].

3. Kalman Filter

KF is an optimal estimator which can result in an optimum estimation of a system state using state space principle and system error modeling. KF is a linear, unbiased and

recursive algorithm that optimally estimates the unknown state of a dynamic system from noisy data taken at discrete real time intervals by minimizing the mean-squared error. The most important feature of KF is that it is recursive, hence it does not require storage of all past observations; the KF algorithm is ideally suited to dealing with complex estimation problem [11].

To achieve recursive equations of KF, we start with state and measurement equations [12]:

$$X_{k+1} = \phi_k X_k + W_k \quad (1)$$

$$Z_k = H_k X_k + V_k \quad (2)$$

where X_k is the $n \times 1$ state vector at time t_k, ϕ_k is the $n \times n$ state transition matrix from X_k to X_{k+1}, $W_k \sim N(0, Q_k)$ is the $n \times 1$ process error vector (a white sequence with a specific covariance-based function is assumed), Z_k is the $m \times 1$ measurement vector at time t_k, H_k is the $m \times n$ matrix which creates an ideal relation (without any noises) between measurement and state vectors at time t_k and $V_k \sim N(0, R_k)$ is the $m \times 1$ error vector (a white sequence is recognized with covariance structure and is supposed as W_k with cross correction equal to zero). The KF recursive equations are:

$$K_k = P_k^- H_k^T (H_k P_k^- H_k^T + R_k)^{-1} \quad (3)$$

where K_k is called Kalman gain. The estimation process is updated as:

$$\hat{X}_k = \hat{X}_k^- + K_k (Z_k - H\hat{X}_k^-) \quad (4)$$

where "^" and "-" denote estimated state and prior to measurement incorporation, respectively. By calculating the corresponding covariance matrix using optimum estimation, we will have:

$$P_k = (I - K_k H_k) P_k^- \quad (5)$$

Consequently, optimum state estimation and covariance matrix are resulted for the next time step.

$$X_{k+1}^- = \phi_k \hat{X}_k \quad (6)$$

$$P_{k+1}^- = \phi_k P_k \phi_k^T + Q_k \quad (7)$$

4. Genetic Algorithm

GA is an effective searching method in a very wide and huge space. It affects getting an optimum result that it is probably not possible to achieve in a person's life. GAs are far more different from primitive optimization methods. In these algorithms, the design space must be changed into the genetic representation. Therefore, GAs deal with a series of encoded variables. The advantage of using encoded variables is that it can be possible to encode continuous functions like discrete functions. GA is based on random processing or more specifically it is based on guided random process. Therefore, random operators of searching space are examined in a comparative way. Basically, in order to use GA these three important concepts must be defined. If these three sections are defined correctly, GA will certainly function properly and it is finally possible to improve the performance of the system by applying some changes.

Objective function: In each problem, the purpose is to maximize or minimize a parameter or parameters. Therefore the objective function is determined using mathematical relations and proper weighing to solve the problem.

Searching space: The purpose of problem solving is to find the best result among different results. The space of all probable states is called searching space. Each result could be represented by a value which determines its propriety. Searching for a result means finding an extremum result within the searching space.

Operators of GA: After achieving the objective function and encoding the population, it is the time for operators of GA to start functioning. In a simple GA, the following three main operators including reproduction, merging and mutation operators are usually used.

Reproduction is usually the first step which is applied to the population. In this method, a series of genes are selected as a parent in the population which will result in generating children by merging. Based on the theory of the fittest, the best individuals should be selected in order to generate the best next generation. Correspondingly, the reproduction operator is sometimes called selecting operator. There are different selection methods in GA to select genes, but the purpose for all of them is to select strings with high average between current population and producing multiple copies of them and putting them in a place, which is called reproduction pool based on a probable form.

After the reproduction stage is complete, a population of the fittests is generated. In fact reproduction function has selected a set of the best strings (colony), but it has not generated new strings. For this reason, merging function is applied to reproduction pool in order to produce new better strings. The purpose of merging is to search the parameter space and to preserve hidden information in strings as much as possible.

Mutation of a bit is to change between 0 and 1 and vice versa bit by bit which is done based on a small probability like P_m. In mutation stage a random number between 0 and 1 is produced. If the produced number is less than P_m then the output is considered as true and otherwise it is considered as false. If the output is true for each bit, the bit will change, otherwise it will remain unchanged.
Bits of a string are mutated independently; it means that mutation of a certain bit does not affect the probability of other bits. In a simple GA, this function is considered as a secondary operator to preserve the information which might be missing. For example, consider that all bits of strings in a population in a certain range are zero and optimum solving method needs one 1 in that point, while merging operator can not generate 1 in that situation. Therefore, in order to generate 1, mutation function is used. In the following sections objective function and the operators which are used in this paper are explained.

5. KF Parameters Tuned using GA

In order to utilize KF, it is necessary to know the accurate mathematical model. However, it is impossible to achieve an accurate model due to lots of uncertainties existing in GPS including errors of the time-variant disturbance, parameter perturbation in mathematical model and unknown statistical properties of noise existing in the system. These factors will result in estimate error in KF, even they make it unstable.
In order to achieve the best filtering performance, the two parameters, that is, system model parameter error variance matrix Q_k and measurement noise sequence variance matrix R_k must be optimized by GA [13]. The objective function J will reach the minimum value by using the combination of best filtering effect. The objective function J is described as follow:

$$J = \sqrt{RMS_x^2 + RMS_y^2 + RMS_z^2} \quad (8)$$

in which RMS_x, RMS_y and RMS_z are effective errors of x, y and z, respectively. The traditional parameter optimization method based on calculations is derived from the assumption that the resolving function is continuous and differentiable, but in practice the continuous and differentiable condition of the function is not fulfilled and results in several solutions and heavy noise.
Compared to traditional methods, GA is not influenced by continuous and differentiable condition, and it can solve many complex problems, which can not be done using traditional methods. It possesses extensive adaptability and good robust performance, and it is easily applicable in parallel tasks. The operators which are used in the implemented GA in this paper are mentioned in the following section:

Step 1: Population: The algorithm starts at the first place with a set of random results, which is called population. These results are used to generate next new population anticipating a better population than the previous ones, because the methods are used for selecting new populations with respect to their propriety. Therefore, the best ones have more chances to reproduce. This process repeats until final conditions are met (to achieve the best results). After doing different experiments, the size of population was decided to contain 165 subjects in this paper.

Step 2: Reproduction: Between available methods to select genes and to merge them the tournament method has been selected. This method is similar to a tournament in the nature, in which a small subset of genes is selected randomly and competes with others. Finally, one of them wins with respect to its propriety in this tournament and then it is copied in the mating pool as a new parent and this process repeats until all parents will be generated in the new population.

Step 3: Crossover: Crossover options specify how the GA combines two individuals, or parents, to form a crossover child for the next generation. In this paper, the heuristic method has been selected after considering different merging methods.

Step 4: Mutation: After merging strings, it is the time to mutate. Mutation is useful for preventing quick convergence and helping the search algorithm to escape from being trapped in positional minimums. On the other hand, this function is used to preserve different states of genes and to be distinct in a population. The mutation which is used in this paper is mutation adapt feasible. The feasible region is bounded by the constraints and inequality constraints. A step length is chosen along each direction so that linear constraints and bounds are satisfied.

Step 5: Migration: Migration options specify how individuals move between subpopulations. Migration occurs if we set population size to be a vector of length greater than 1. When migration occurs, the best individuals from one subpopulation replace the worst individuals in another subpopulation. Individuals that migrate from one subpopulation to another are copied. They are not removed from the source subpopulation. Migration wraps at the ends of the subpopulations. That is, the last subpopulation migrates into the first, and the first may migrate into the last.

Step 6: Stopping criteria: Stopping criteria determine what causes the algorithm to terminate. We can specify the following options:

- **Function generations:** Specifies the maximum number of iterations for the GA to perform. The adjusted value in this paper is 80.
- **Function stall generations:** The algorithm stops if the weighted average change in the fitness function value over stall generations is less than the function tolerance. The adjusted value in this paper is 70.
- **Function stall time limit:** The algorithm stops if there is no improvement in the best fitness value for an interval of time in seconds specified by stall time. The adjusted value in this paper is 2000 seconds.

6. Experimental Results

Data collection is highly important in assessing the utmost performance and efficiency of the method suggested in this paper. Such a process was carried out in the center for Computer Control and Fuzzy Logic Research Lab in Iran University of Science and Technology. Fig. 1 depicts the hardware used in data collection process.
According to the hardware shown in Fig. 1, the serial GPS receiver data are passed through TTL-RS232 converter to change their levels from TTL to RS232 standard and make it ready to be connected to the computer. It should be noted that the keyboard on this hardware board is used for the purpose of setting GPS receiver's programmable parameters such as, the output protocol of receiver's serial ports (NMEA or Binary) and data transmission rate (4600 or 9600 bit/s) [14].

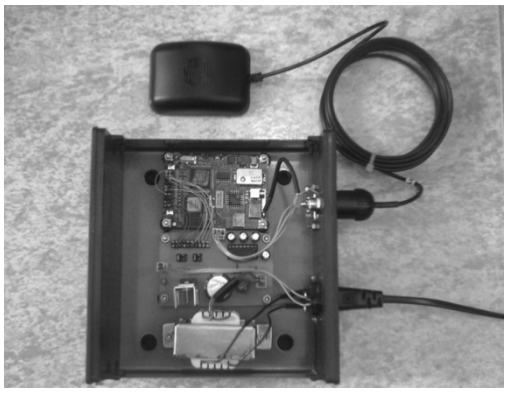


Fig. 1: GPS receiver board

The technically significant features of the GPS receiver used in data collection process include 5-channel GPS receiver, capable of keeping track of up to 9 satellites, capable of reducing SA effect in static mode, functioning with both active and inactive antennas, capable of position measuring with maximum accuracy in SPS mode, capable of selecting satellites and making satellite's view angle narrow and capable of updating information in each second.
Considering the precision and the convergence rate, filtering result J is adopted as an objective function. Two parameter matrixes, system model parameter error variance matrix Q_k , and measurement noise sequence variance matrix R_k are regarded as the optimal parameters for KF which make J reach the minimum value. Table 1 reveals the features of the GA which is utilized in this paper.

Table 1: Features of the GA utilized in this paper

Parameter	Features
Population size	165
Generations	80
Stall gen limit	70
Stall time limit	2000
Initial population	R1,R2,R3
Mutation function	Adapt feasible
Selection function	Tournament
Crossover function	Heuristic

The optimal values of Q_k and R_k were put in KF relation and the filtering experiment was carried out on 1000 data which were obtained from the aforementioned GPS. Figures 2 to 5 represent the real, predicted and prediction error values of the component position errors for 1000 experimental data sets using KF tuned by GA in SA error turned on and turned off states.

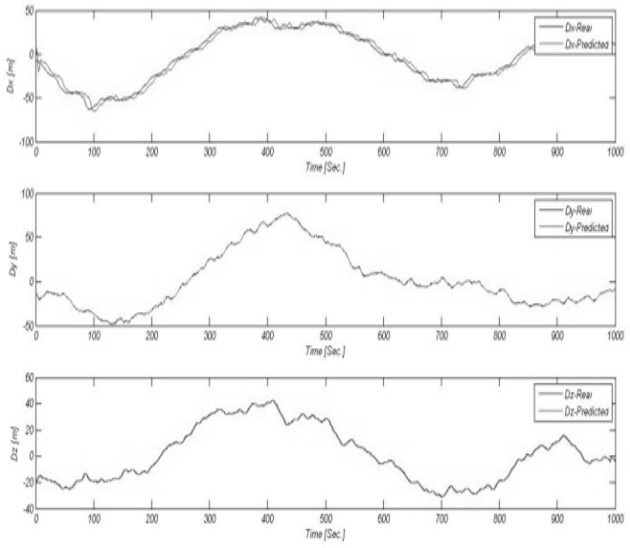


Fig. 2: The results of 1000 prediction for component positions using KF tuned by GA (SA on)

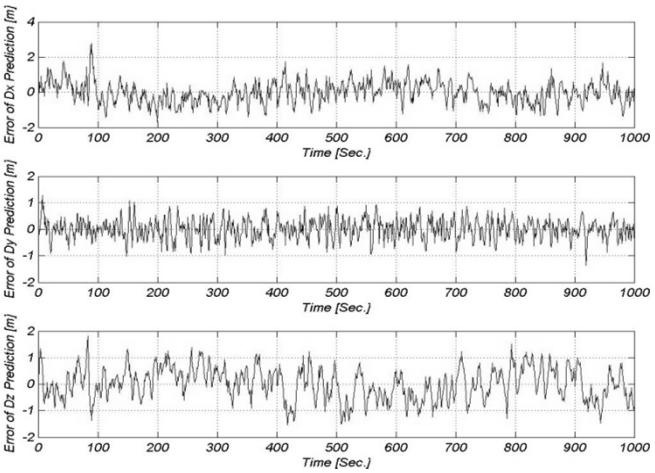


Fig. 3: The results of 1000 prediction error for component positions using KF tuned by GA (SA on)

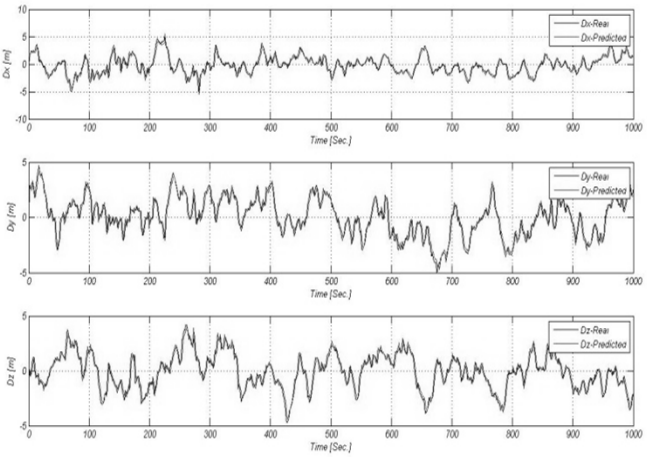


Fig. 4: The results of 1000 prediction error for component positions using KF tuned by GA (SA off)

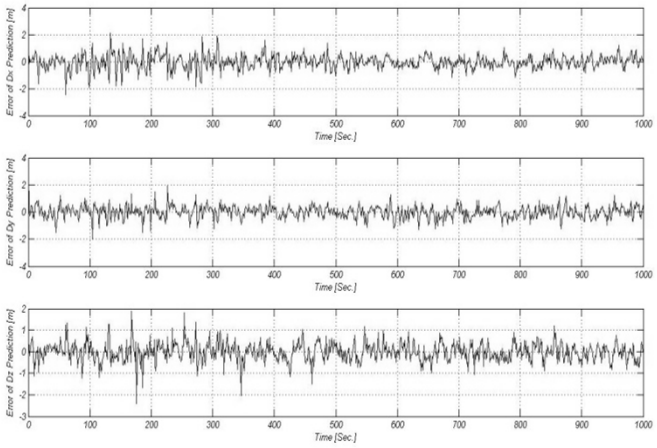


Fig. 5: The results of 1000 prediction error for component positions using KF tuned by GA (SA off)

Tables 2 and 3 depict the statistical features of estimation errors for 1000 tests which were performed on experimental data. According to the results illustrated in Tables 2 and 3, it is noticeable that the RMS errors in estimation error in component positions using KF tuned by GA in SA on and off modes reduced to less than 1 and 0.9 meters, respectively. The results from tests carried out on real data show that the functionality of KF tuned by GA in estimating components of position errors is independent of the effect of SA errors which is one of the advantages of KF tuned by GA.

References [15,16] predict DGPS corrections using traditional KF. As shown in Table 4, the KF tuned by GA has better accuracy than traditional KF for DGPS corrections prediction. There are few papers that predict the DGPS corrections using KF. The proposed KFs in this paper have more accuracy than them.

Table 2: The maximum, minimum, average and RMS error values for 1000 estimation using KF adapted with GA (SA off)

Parameters	X	Y	Z
Maximum	2.1349	3.0893	1.8759
Minimum	-2.4570	-2.0089	-2.4190
Average	0.0145	-0.0091	0.0045
RMS	0.5170	0.4546	0.4256
Total RMS	**0.8093**		

Table 3: The maximum, minimum, average and RMS error values for 1000 estimation using KF adapted with GA (SA on)

Parameters	X	Y	Z
Maximum	2.7732	1.2690	1.8113
Minimum	-1.7061	-2.4190	-1.5343
Average	0.0188	-0.0048	-0.0221
RMS	0.6268	0.3815	0.6044
Total RMS	**0.9507**		

Table 4: Comparing DGPS corrections prediction accuracy using proposed method and traditional KF

Prediction Method	Accuracy (SA on)	Accuracy (SA off)
Traditional KF [15,16]	1.4380	0.8108
KF tuned by GA	0.9507	0.8093

7. Conclusions

Regarding the increasing development in using GPS and their pivotal roles in various fields including business and the military, improvement in their measurement accuracy and data security in these systems are considered not only as theoretical issues, but also as vital necessities in these systems. In this paper, the way of utilizing a low cost GPS receiver as an accurate positioning device was discussed.

Moreover, a KF-based GA was proposed. The experimental results using real data which were gathered in test fields, affirm the high potential of these methods to obtain precise positioning information. The results reveal that it is possible to reduce position RMS error in single-frequency GPS receivers to less than 1 meter, especially when SA is in on mode, the effective error value improves to 0.4873 meter utilizing KF adapted with GA as compared to traditional KF.

References

[1] K. D. McDonald, "The Modernization of GPS: Plans, New Capabilities and the Future Relationship to Galileo", Journal of Global Positioning System, Vol.1, No.1, pp.1-17, 2002.

[2] M. R. Mosavi, "GPS Receivers Timing Data Processing using Neural Networks: Optimal Estimation and Errors Modeling", Journal of Neural Systems, World Scientific, Vol.17, No.5, pp.383-393, 2007.

[3] P. Misra, B. P. Burke and M. M. Pratt, "GPS Performance in Navigation", Proceeding of the IEEE, Vol.87, No.1, pp.65-85, 1999.

[4] K. Kobayashi, Ka C. Cheok, K. Watanabe and F. Munekata, "Accurate Differential Global Positioning System via Fuzzy Logic Kalman Filter Sensor Fusion Technique", IEEE Transactions on Industrial Electronics, Vol.45, No.3, pp.510-518, 1998.

[5] A. Indriyatmoko, T. Kang, Y. J. Lee, G. I. Jee, Y. B. Cho and J. Kim, "Artificial Neural Networks for Predicting DGPS Carrier Phase and Pseudo-Range Correction", Journal of GPS Solutions, Vol.12, No.4, pp.237-247, 2008.

[6] B. W. Parkinson, J. J. Spilker Jr, P. Axelrad and P. Enge, "Global Positioning System: Theory and Applications", The American Institute of Aeronautics and Astronautics, 1996.

[7] R. Yarlagadda, I. Ali, N. Al-Dhahir and J. Hershey, "GPS GDOP Metric", IEE Proc.-Radar, Sonar Navig., Vol.147, No.5, pp.259-264, 2000.

[8] O. Øvstedal, "Absolute Positioning with Single-Frequency GPS Receivers", Journal of GPS Solutions, Vol.5, No.4, pp.33-44, 2002.

[9] G. Seeber, F. Menge and C. Volksen, "Precise GPS Positioning Improvements by Reducing Antenna and Site Dependent Effects", Advances in Positioning and Reference Frames, IAG Symposium, Vol.118, pp.237-244, 1997.

[10] K. Deergha Rao, "An Approach for Accurate GPS Navigation with SA", IEEE Transactions on Aerospace and Electronic Systems, Vol.34, No.2, pp.695-699, 1998.

[11] D. Simon, "Optimal State Estimation: Kalman, H. Infinity and Nonlinear Approches", John Wiley, 2006.

[12] W. S. Chaer, R. H. Bishop and J. Ghosh, "A Mixture-of-Experts Framework for Adaptive Kalman Filtering", IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics, Vol.27, No.3, pp.452-464, 1997.

[13] J. Yan, D. Yuan, X. Xing and Q. Jia, "Kalman Filtering Parameter Optimization Techniques Based on Genetic Algorithm", IEEE Conference on Automation and Logistics, pp.1717-1720, 2008.

[14] "Zodiac GPS Receiver Family Designer's Guide", Rockwell Semiconductor Systems, GPS-33, 1996.

[15] M. R. Mosavi, "Comparing DGPS Corrections Prediction using Neural Network, Fuzzy Neural Network and Kalman Filter", Journal of GPS Solutions, Vol.10, No.2, pp.97-107, 2006.

[16] M. R. Mosavi, A. Nakhaei and Sh. Bagherinia, "Improvement in Differential GPS Accuracy using Kalman Filter", Journal of Aerospace Science and Technology, Vol.7, No.2, pp.139-150, 2010.

Mohammad-Reza Mosavi (Corresponding Author) received his B.S., M.S., and Ph.D. degrees in Electronic Engineering from Department of Electrical Engineering, Iran University of Science and Technology (IUST), Tehran, Iran in 1997, 1998, and 2004, respectively. He is currently faculty member of Department of Electrical Engineering of IUST as associate professor. He is the author of about 120 scientific publications on journals and international conferences. His research interests include Artificial Intelligent Systems, Global Positioning Systems, Geographic Information Systems and Remote Sensing.

Performance Analysis of Enhanced Clustering Algorithm for Gene Expression Data

T.Chandrasekhar[1], K.Thangavel[2] and E.Elayaraja[3]

**[1] Research Scholar, Bharathiar university,
Coimbatore, Tamilnadu, India - 641 046.**

**[2] Department of Computer Science, Periyar University,
Salem, Tamilnadu, India -636 011.**

**[3] Department of Computer Science, Periyar University,
Salem, Tamilnadu, India -636 011.**

Abstract

Microarrays are made it possible to simultaneously monitor the expression profiles of thousands of genes under various experimental conditions. It is used to identify the co-expressed genes in specific cells or tissues that are actively used to make proteins. This method is used to analysis the gene expression, an important task in bioinformatics research. Cluster analysis of gene expression data has proved to be a useful tool for identifying co-expressed genes, biologically relevant groupings of genes and samples. In this paper we applied K-Means with Automatic Generations of Merge Factor for ISODATA- AGMFI. Though AGMFI has been applied for clustering of Gene Expression Data, this proposed Enhanced Automatic Generations of Merge Factor for ISODATA- EAGMFI Algorithms overcome the drawbacks of AGMFI in terms of specifying the optimal number of clusters and initialization of good cluster centroids. Experimental results on Gene Expression Data show that the proposed EAGMFI algorithms could identify compact clusters with perform well in terms of the Silhouette Coefficients cluster measure.

Keywords: *Clustering, K-Means, AGMFI, EAGMFI, Gene expression data.*

1. Introduction

Clustering has been used in a number of applications such as engineering, biology, medicine and data mining. Cluster analysis of gene expression data has proved to be a useful tool for identifying co-expressed genes. DNA microarrays are emerged as the leading technology to measure gene expression levels primarily, because of their high throughput. Results from these experiments are usually presented in the form of a data matrix in which rows represent genes and columns represent conditions [12]. Each entry in the matrix is a measure of the expression level of a particular gene under a specific condition. Analysis of these data sets reveals genes of unknown functions and the discovery of functional relationships between genes [18]. The most popular clustering algorithms in microarray gene expression analysis are Hierarchical clustering [11], K-Means clustering [3], and SOM [8]. Of these K-Means clustering is very simple and fast efficient. This is most popular one and it is developed by Mac Queen [6]. The easiness of K-Means clustering algorithm made this algorithm used in several fields. The K-Means algorithm is effective in producing clusters for many practical applications, but the computational complexity of the original K-Means algorithm is very high, especially for large data sets. The K-Means clustering algorithm is a partitioning clustering method that separates the data into K groups. One drawback in the K-Means algorithm is that of a priori fixation of number of clusters [2, 3, 4, 17].

Iterative Self-Organizing Data Analysis Techniques (ISODATA) tries to find the best cluster centres through iterative approach, until some convergence criteria are met. One significant feature of ISODATA over K-Means is that the initial number of clusters may be merged or split, and so the final number of clusters may be different from the number of clusters specified as part of the input. The ISODATA requires number of clusters, and a number of additional user-supplied parameters as inputs. To get better results user need to initialize these parameters with appropriate values by analyzing the input microarray data. In [10] Karteeka Pavan et al proposed AGMFI algorithm to initialize merge factor for ISODATA. This paper studies an initialization of centroids proposed in [17] for microarray data to get the best quality of clusters. A comparative analysis is performed for UCI data sets in-order to get the best clustering algorithm.

This paper is organised as follows. Section 2 presents an overview of Existing works K-Means algorithm, Automatic Generation of Merge Factor for ISODATA (AGMFI) methods. Section 3 describes the Enhanced initialization algorithm. Section 4 describes performance study of the above methods for UCI data sets. Section 5 describes the conclusion and future work.

2. Related Work

2.1 K- Means Clustering

The main objective in cluster analysis is to group objects that are similar in one cluster and separate objects that are dissimilar by assigning them to different clusters. One of the most popular clustering methods is K-Means clustering algorithm [3, 9, 12, 17]. It classifies object to a pre-defined number of clusters, which is given by the user (assume *K* clusters). The idea is to choose random cluster centres, one for each cluster. These centres are preferred to be as far as possible from each other. In this algorithm mostly Euclidean distance is used to find distance between data points and centroids [6]. The Euclidean distance between two multi-dimensional data points $X = (x_1, x_2, x_3, ..., x_m)$ and $Y = (y_1, y_2, y_3, ..., y_m)$ is described as follows:

$$D(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_m - y_m)^2}$$

The K-Means method aims to minimize the sum of squared distances between all points and the cluster centre. This procedure consists of the following steps, as described below.

Algorithm 1: K-Means clustering algorithm [17]

Require: D = {d_1, d_2, d_3, ..., d_n } // Set of n data points.
K - Number of desired clusters
Ensure: A set of K clusters.

Steps:
1. Arbitrarily choose *k* data points from *D* as initial centroids;
2. **Repeat**
   Assign each point d_i to the cluster which has the closest centroid;
   Calculate the new mean for each cluster;
   Until convergence criteria is met.

Though the K-Means algorithm is simple, it has some drawbacks of quality of the final clustering, since it highly depends on the arbitrary selection of the initial centroids [1].

2.2 Automatic Generation of Merge Factor for ISODATA (AGMFI) Algorithm

The clusters produced in the K-Means clustering are further optimized by ISODATA algorithm. Some of the parameters are fixed by user during the merging and partitioning the clusters. In [10], Automatic Generation of Merge Factor is proposed to initialize merge factor for ISODATA. AGMFI uses different heuristics to determine when to split. Decision of merging is done based upon merge factor which is the function of distances between the clusters. The step by step procedure of AGMFI is given here under.

Algorithm 3: The AGMFI algorithm [10]

Require: D = {d_1, d_2, d_3, ..., d_n } // Set of n data points.
K - Number of desired clusters.
m- minimum number of samples in a cluster.
N – maximum number of iterations.
Θs – a threshold value for spilt_size.
Θc - a threshold value for merge_size.
Ensure: A set of K clusters.

Steps:
1. Identify clusters using K-Means algorithms;
2. Find the inter distance in all other cluster to minimum average inter distances clusters point in C;
3. Discard the m and merging operations of cluster ≥ 2*K, If n is even go to step 4 or 5;
4. Distance between two centroids < C, merge the Cluster and update centroid, otherwise repeat up to K/2 times;
5. K ≤ K/2 or n is odd go to step 6 or 7;
6. Find the standard division of all clusters that has exceeds S * standard division of D;
7. Executed N times or no changes occurred in clusters since the last time then stop, otherwise take the centroids of the clusters as new seed points and find the clusters using K-Means and go to step 3.

The main difference between AGMFI and ISODATA is ISODATA uses heuristic values to merge the clusters, AGMFI generates automatically and the choice of c is not fixed but is to be decided to have better performance. The distance measure used here is the Euclidean distance. To assess the quality of the clusters, we used the silhouette measure proposed by Rousseeuw [14].

3. The Enhanced Method

Performance of iterative clustering algorithms which converges to numerous local minima depends highly on initial cluster centers. Generally initial cluster centers are selected randomly. In this section, the cluster centre initialization algorithm is studied to improve the performance of the K-Means algorithm.

Algorithm 2: The Enhanced Method [17]

Require: $D = \{d_1, d_2, d_3,..., d_i,..., d_n\}$ // Set of n data points.
$d_i = \{x_1, x_2, x_3,..., x_i,..., x_m\}$ // Set of attributes of one data point.
k // Number of desired clusters.
Ensure: A set of k clusters.

Steps:
1. In the given data set D, if the data points contains the both positive and negative attribute values then go to Step 2, otherwise go to step 4.
2. Find the minimum attribute value in the given data set D.
3. For each data point attribute, subtract with the minimum attribute value.
4. For each data point calculate the distance from origin.
5. Sort the distances obtained in step 4. Sort the data points accordance with the distances.
6. Partition the sorted data points into k equal sets.
7. In each set, take the middle point as the initial centroid.
8. Compute the distance between each data point d_i $(1 \le i \le n)$ to all the initial centroids c_j $(1 \le j \le k)$.
9. **Repeat**
10. For each data point d_i, find the closest centroid c_j and assign d_i to cluster j.
11. Set ClusterId[i]=j. // j:Id of the closest cluster.
12. Set NearestDist[i]= $d(d_i, c_j)$.
13. For each cluster j $(1 \le j \le k)$, recalculate the centroids.
14. **For** each data point d_i,
14.1 Compute its distance from the centroid of the present nearest cluster.
14.2 If this distance is less than or equal to the present nearest distance, the data point stays in the same cluster.
Else
14.2.1 For every centroid c_j $(1 \le j \le k)$ compute the distance $d(d_i, c_j)$.
End for;
Until the convergence criteria is met.

4. Experimental Analysis and Discussion

The following data sets are used to analyse the methods studied in sections 2 and 3.

Serum data
This data set is described and used in [10]. It can be downloaded from: http://www.sciencemag.org/feature/data /984559.shl and corresponds to the selection of 517 genes whose expression varies in response to serum concentration inhuman fibroblasts.

Yeast data
This data set is downloaded from Gene Expression Omnibus-databases. The Yeast cell cycle dataset contains 2884 genes and 17 conditions. To avoid distortion or biases arising from the presence of missing values in the data matrix we removal all the genes that had any missing value. This step results in a matrix of size 2882 * 17.

Simulated data
It is downloaded from http://www.igbmc.ustrasbg.fr/projets/fcm/y3c.txt. The set contains 300 Genes [3]. Above the microarray data set values are all normalized in every gene average values zero and standard deviation equal to 1.

4.1 Comparative Analysis

The K-Means, Enhanced with K-Means and AGMFI are applied on serum data set when number of clusters is taken as 10 and 5 times running to EAGMFI clusters data into 7.

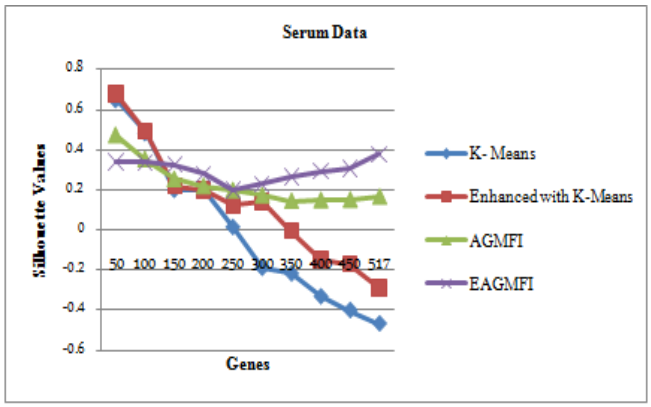


Fig. 1 Performance Comparison chart for serum data.

K-Means, Enhanced with K-Means and AGMFI are applied on Yeast data set when number of clusters initialized to 34 and 5 times running on EAGMFI clusters data into 18 .

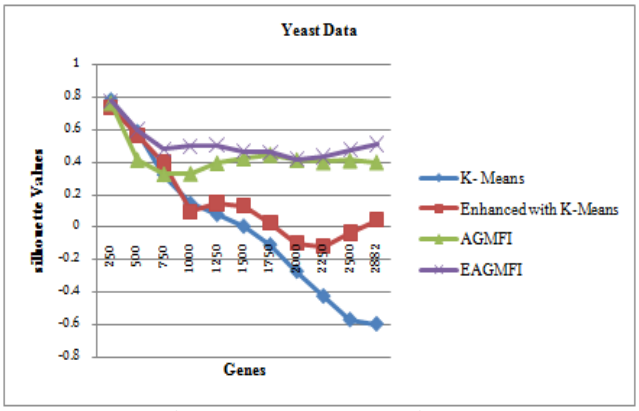


Fig. 2 Performance Comparison chart for Yeast data.

The K-Means, Enhanced with K-Means and AGMFI are applied on simulated data set when number of clusters initialized to 10 and 5 times running to EAGMFI clusters data into 6.

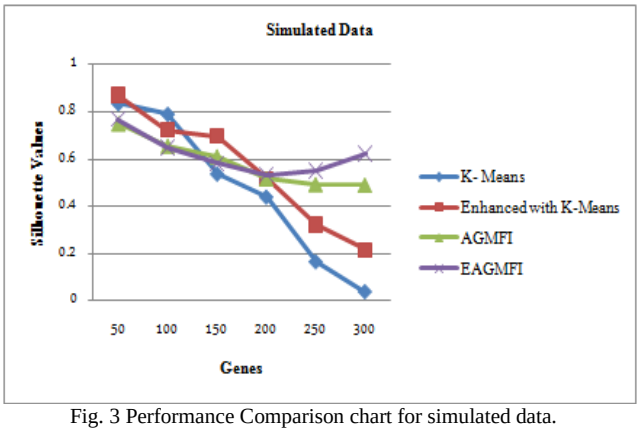


Fig. 3 Performance Comparison chart for simulated data.

Table 1: Comparative Analysis of Clustering Quality of Measurement.

Data set	Initial No of cluster	Finalized No of cluster	Cluster Quality by K-Means	Cluster Quality by Enhanced with K-Means	Cluster Quality by AGMFI	Cluster Quality by EAGMFI
Serum	10	7	-0.566	12.38	22.78	29.594
Yeast	34	18	-0.43	17.27	43.26	51.38
simulated	10	6	46.74	55.79	58.695	61.79

It is observed from the above analysis that the proposed method is performing well for all the three data cells.

5. Conclusion

In this paper AGMFI was studied to improve the quality of clusters. The Enhanced Automatic Generation of Merge Factor for ISODATA (EAGMFI) Clustering Microarray Data based on K-Means and AGMFI clustering algorithms were also studied. One of the demerits of AGMFI is random selection of initial seed point of desired clusters. This was overcome with Enhanced for finding the initial centroids algorithms to avoidance for initial values at random. Therefore, the EAGMFI algorithm not depending upon the any choice of the number of cluster and automatic evaluation initial seed of centroids it produces different better results with Silhouette Coefficients measurement. Both the algorithms were tested with gene expression data.

References

[1] A. M. Fahim, A. M. Salem, F. A. Torkey and M. A. Ramadan, "An Efficient enhanced K-Means clustering algorithm", *journal of Zhejiang University*, 10 (7): 1626 - 1633, 2006.

[2] Bashar Al-Shboul and Sung-Hyon Myaeng, "Initializing K-Means using Genetic Algorithms", *World Academy of Science, Engineering and Technology* 54, 2009.

[3] Chen Zhang and Shixiong Xia, " K-Means Clustering Algorithm with Improved Initial center," *in Second International Workshop on Knowledge Discovery and Data Mining* (WKDD), pp. 790-792, 2009.

[4] F. Yuan, Z. H. Meng, H. X. Zhangz, C. R. Dong, " A New Algorithm to Get the Initial Centroids", *proceedings of the 3rd International Conference on Machine Learning and Cybernetics*, pp. 26-29, August 2004.

[5] Chaturvedi J. C. A, Green P, "K - Modes clustering," *Journals of Classification*, (18):35–55, 2001.

[6] Doulaye Dembele and Philippe Kastner, "Fuzzy C means method for clustering microarray data", *Bioinformatics*, vol.19, no.8, pp.973- 980, 2003.

[7] Daxin Jiang, Jian Pei, and Aidong Zhang "An Interactive Approach to mining Gene Expression Data". *IEEE Transactions on Knowledge and Data Engineering*, vol 17, No.10, pp.1363- 1380, October 2005.

[8] Dongxiao Zhu, Alfred O Hero, Hong Cheng, Ritu Khanna and Anand Swaroop, "Network constrained clustering for gene microarray Data", *doi*:10.1093 /*bioinformatics* / *bti* 655, Vol. 21 no. 21, pp. 4014 – 4020, 2005.

[9] Fahim A.M, Salem A. M, Torkey A and Ramadan M. A, "An Efficient enhanced K-Means clustering algorithm", *Journal of Zhejiang University*, 10(7):1626–1633, 2006.

[10] K Karteeka Pavan, Allam Appa Rao, A V Dattatreya Rao, GR Sridhar, "Automatic Generation of Merge Factor for Clustering Microarray Data", *IJCSNS International Journal of Computer Science and Network Security*, Vol.8, No.9, September 2008.

[11] Kohei Arai and Ali Ridho Barakbah, " Hierarchical K-Means: an algorithm for centroids initialization for K-Means", *Reports of the Faculty of Science and Engineering, Saga University*, Vol. 36, No.1, 25-31, 2007.

[12] K.R De and A. Bhattacharya , "Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying Patterns in expression profiles," *bioinformatics*, Vol. 24, pp.1359- 1366, 2008.

[13] K. A. Abdul Nazeer and M. P. Sebastian, "Improving the accuracy and efficiency of the K-Means clustering algorithm", in *nternational Conference on Data Mining and Knowledge Engineering (ICDMKE),Proceedings of the World Congress on Engineering (WCE-2009), London, UK.* Vol 1, July 2009.

[14] Lletí, R., Ortiz, M.C., Sarabia, L.A., Sánchez, M.S. "Selecting variables for K-Means cluster analysis by using a genetic algorithm that optimises the silhouettes". *Analytica Chimica Acta*, 2004.

[15] Moh'd Belal Al- Zoubi and Mohammad al Rawi, "An Efficient Approach for Computing Silhouette Coefficients". *Journal of Computer Science* 4 (3): 252-255, 2008.

[16] Margaret H. Dunham, "Data Mining- Introductory and Advanced Concepts", *Pearson education*, 2006.

[17] Madhu Yedla, Srinivasa Rao Pathakota, T M Srinivasa , "Enhancing K-Means Clustering Algorithm with Improved Initial Center", Madhu Yedla et al. / *(IJCSIT) International Journal of Computer Science and Information Technologies*, Vol. 1 (2), pp121-125, 2010.

[18] Sunnyvale, Schena M. " Microarray biochip technology ". *CA: Eaton Publishing*; 2000.

[19] Wei Zhong, Gulsah Altun, Robert Harrison, Phang C. Tai, and Yi Pan, "Improved K-Means Clustering Algorithm for Exploring Local Protein Sequence Motifs Representing Common Structural Property", *IEEE transactions on nanobioscience*, vol. 4, no. 3, september 2005.

[20] Y. Lu, S. Lu, F. Fotouhi, Y. Deng, and S. Brown, "Incremental Genetic K-Means Algorithm and its Application in Gene Expression Data Analysis", BMC Bioinformatics, 2004.

Phishing Attack Protection (PAP) Approaches for Fairness in Web Usage

Mohiuddin Ahmed[1], Jonayed Kaysar[2]

[1] Department of Computer Science and Information Technology, Islamic University of Technology Board Bazar, Gazipur-1704, Bangladesh

[2] Department of Computer Science and Information Technology, Islamic University of Technology Board Bazar, Gazipur-1704, Bangladesh

Abstract

Phishing scams are considered as a threat issue to all web users. But still the web users are not consciously aware of this fact. Many research works have been done to increase the phishing awareness among the users but it is not up to the mark till to date. We have conducted a survey among a diversified group of people who are active user of internet. And then analyzed the existing phishing warnings provided by the web browsers and protection schemes, in this paper we have suggested new approaches i.e. sending notifications to user, checking URL, creating user alarms and security knowledge to ensure fairness in web usage.

Keywords: *Phishing, Design, Warnings, Usable Privacy & Security, Spoof, Phisher.*

1. Introduction

Phishing is an e-mail fraud method in which the perpetrator sends out legitimate-looking email in an attempt to gather personal and financial information from recipients. Typically, the messages appear to come from well known and trustworthy web sites. Web sites that are frequently spoofed by phishers include PayPal, eBay, MSN, Yahoo, BestBuy and America Online[9]. A phishing expedition, like the fishing expedition it's named for, is a speculative venture: the phisher puts the lure hoping to fool at least a few of the prey that encounter the bait.

Suppose you check your e-mail one day and find a message from your bank. You've gotten e-mail from them before, but this one seems suspicious, especially since it threatens to close your account if you don't reply immediately. What do you do? This message and others like it are examples of phishing, a method of online identity theft. In addition to stealing personal and financial data, phishers can infect computers with viruses and convince people to participate unwittingly in money laundering.

2. A Statistics of Scams

In 2010, RSA[3] witnesses a total of 203,985 phishing attacks launched. As compared to the total in 2009, this marks a 27 percent increase in phishing attack volume over the past year.

Top Ten Countries by Attack Volume

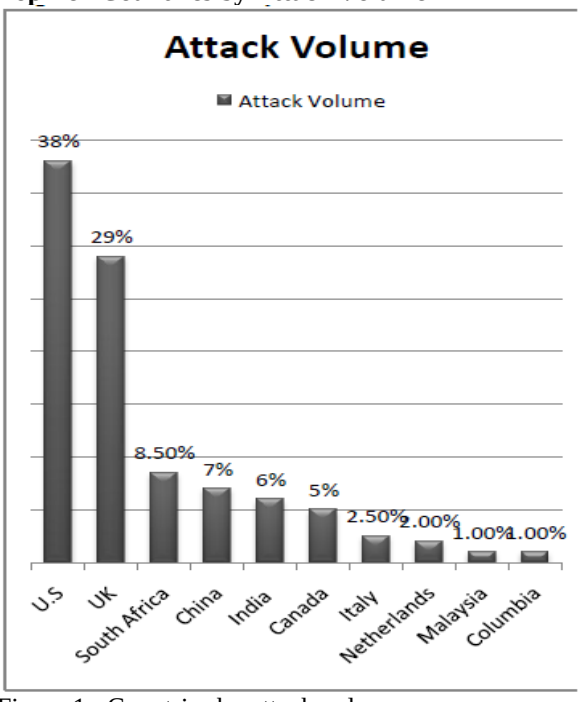


Figure 1 : Countries by attack volume

3. State-of-the-art Web Browsers

The state-of-the-art web browsers have included active warnings, which coerce users to notice the warnings by interrupting. We will consider most favorite browsers i.e. Microsoft's Internet Explorer and Mozilla Firefox. IE 9 includes both active and passive warnings. Upon observing a confirmed phishing site the browser will display an active warning message in full screen with URL bar colored red[10]. For passive indication IE 9 will show a popup dialog box while sensing the site as suspicious[10].

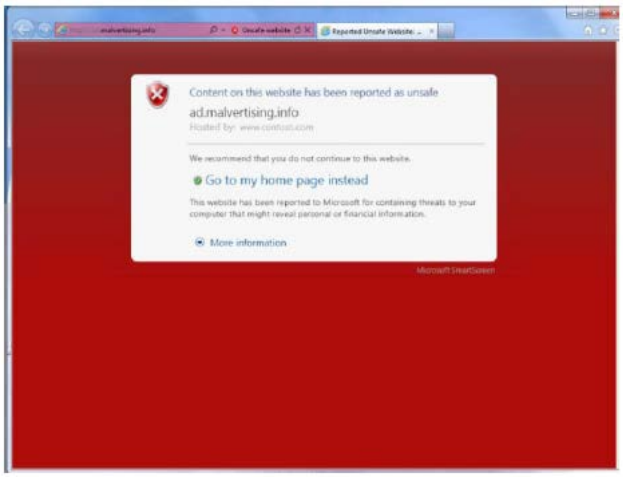


Figure 2 : The active Internet Explorer 9.0 phishing warning[10]

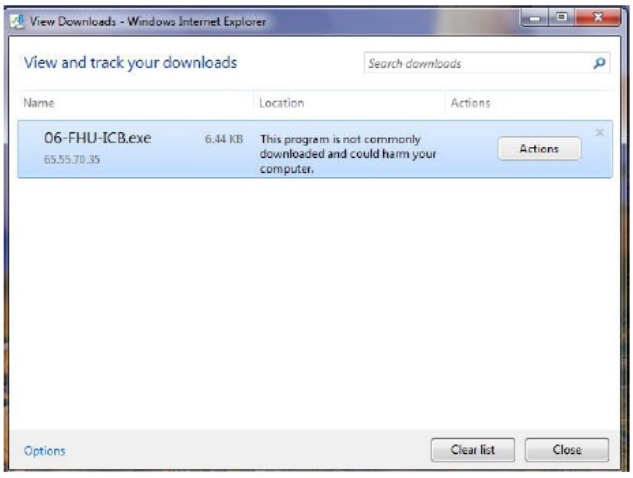


Figure 3 : The passive Internet Explorer 9.0 phishing warning[10]

Firefox 6.0 also includes active cautionary. When a user confronts a phishing site, a non-interactive dimmed version of the site is shown with a dialog box given choices for continuing or leaving the site[11].

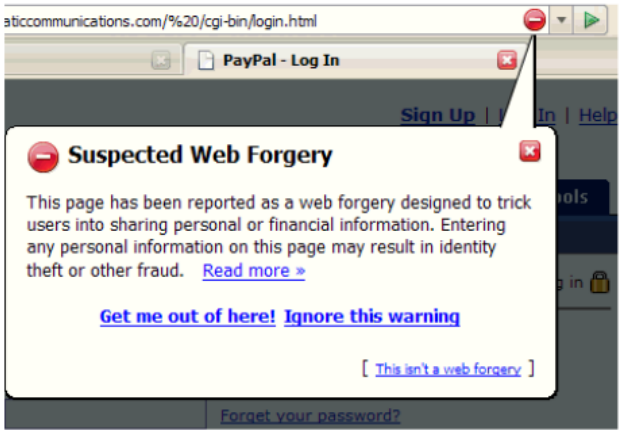


Figure 4 : The active Firefox 6.0 phishing warning[11]

This paper has two contributions. At first, the phishing awareness level of web users is surveyed. And then analyzing the stat-of-the-art web browsers phishing warnings and other approaches it recommends phishing attack protection (PAP) approaches to help users handling the online frauds.

4. Related Works

There have been extensive surveys about phishing awareness to identify the user knowledge on phishing. In [1] the work has simulated a spear phishing attack to expose users to browser warnings. But interestingly enough 97% of the participants fell for at least one of the phishing messages sent to them. And then they used active and passive warnings to find out how the users reacting on that situation. Basically they have used a model form warning science to identify the user reaction and offered suggestion for making more efficient phishing warnings. Another study in [2] shows that the users after having training are less likely to fell for the attacks. The study evaluated the effectiveness of PhishGuru training in field trails and found that generic training materials are more effective than spear phishing training materials.

Even the unaware users of web can be helped not to be in a trap by Client Side Phishing Protection[4]. In this paper they have provided two approaches by which user will be able to identify whether phishing attack is there or not. Their work focused on browser extension which has a limited range and valid for online-banking.

Another approach was based on visual cue. The solution provided by Dhamija et al.[5] consists of dynamic security

skin in the web browser. Here a remote server has to prove its identity in a way that is efficient for human users to examine and cumbersome for phishers. The problem with this approach is that the users must be educated enough about phishing scams.

We considered the last approach in our related work is flow of information based approach. PwdHash[6,7] has been considered as a favorite anti-phishing solution. Which creates passwords for domains like, a password for www.yahoomail.com will be different if it is input to www.attacker.com. Comparing this approach with AntiPhish[8], which keeps eye on sensitive information. Whenever classified information is used in mistrusted websites then a warning is generated and consequent operations are canceled, that means the user should be always vigilant about the information being input in various websites.

5. Questionnaire

We used a set of questions to know that how much knowledge web users are possessing and based on that we work toward designing new approaches. We collected feedback from one hundred users to identify the specific arena where to emphasize for developing new phishing attack protection approaches.

- **What is your internet usage per day?**

We wanted to know how much time a user spends on the web to measure how vulnerable one individual is.

- **Do you prefer online shopping?**

Most of the phishing attacks are placed on online shopping and credit card transaction based. So, this question helped us understanding the scope of a user being cheated.

- **What is your favorite browser?**

Browsers are also creating protection approaches where a user having no idea about phishing might help him/her from being ripped off.

- **Phishing awareness level**

We extracted the level of knowledge about phishing in this question to judge the real life scenario of web users.

- **Do you feel insecure to use internet for phishing attack?**

This question let us know how much a user is worried about phishing attacks and whether users feel insecure over web space or not.

- **Is your browser able to defense phishing attack?**

This tells us whether user is known about phishing or not and whether his/her browser is able to notify any phishing.

- **Did you face any phishing attack?**

This is to know how experienced a user is about being jockeyed.

- **Do you know someone who has financial loss due to phishing attack?**

Awareness of financial losses claimed the information about phishing attacks are disseminated among peer groups and thus being aware of it.

6. PAP Approaches

Here we are suggesting some approaches considering the user knowledge on phishing and the state-of-the-art web browsers. Our proposed approaches will help the users to be more aware of phishing and thus ensure fairness in web usage.

6.1 Sending Notification to User

When a mail or any message from bank or any other financial organization or any credit card payment gateway is sent to the client that organization needs to send a notification to the client's cell phone as a sms. In this era of smart phones, sending notification by sms is the most simple and reliable solution. For this reason, the client phone number should be synchronized with the organization. Whenever there is a mail sent from the organization, the automated system sends a sms momentarily. This will help the clients to understand that the mail is generated by the original organization. And for attackers, it is not possible to send automated notification to the targeted victim's cell phone.

6.2 Checking the URL

An extension can be developed which will check the URL on behalf of the user. It will check the URL and search for the matches of the URL with default search engine of the browser. If the URL matches, then it checks the domains

and sub domains. Now if the sub domain and other part of the URL matches with another domain, then it will generate a message. Most of the time, sub domain is used to create attack. So, if there is a notice, user will check the sub domain itself.

6.3 Creating User Alarm

There should be database of the phishing sites and browsers should always sync with the database. Currently, an alert is given if a site contains malware and same alerts can be created for those phishing sites. When there is an alert, user have to be cautious about the site.

6.4 Check the Redirection

Most of the users are not aware about the redirection. Although browsers like Firefox have the option to show an alert during redirection, nobody cares. This is a vulnerable thing. Redirection should be checked and if the redirected site is not trustworthy, then an alert will also appear and closes the window as well as adding the site with the central phishing database.

6.5 Increase Security Knowledge

The users are the victim and from our survey we found that many of them have heard the term "phishing" but they don't have any clear concept about it. This situation can be improved by increasing security knowledge. Financial organizations like Banks, Insurance companies can organize some weekly/monthly event for their clients where the latest news and views of internet security will be presented. Beside these events, there must be some tutorial(written/video) with the browsers which will teach the users about phishing. As many people will not find enough time to check that tutorial, it can be made mandatory while installing the browsers to watch the tutorials first.

7. Conclusion

Phishing attacks are causing huge financial losses to individual and group. A number of professional and academic protection approaches have been proposed to date. But still there are very less improvement. So, it's the only way when the users are cautious about this issue and thus oriented approaches provided by us in this paper can help to block this burgeoning problem in today's IT world. In our future work, we will be implementing more novel approaches which will supersede the existing approaches for combating phishing attack.

References

[1] Serge Egelman, Lorrie Faith Cranor, Jason Hong. You 've Been Warned: An empirical Study of the Effectiveness of Web Browser Phishing Warnings. CHI 2008,April 5-10,2008,Florence,Italy.
[2] Lesson From a Real World Evaluation of Anti-Phishing Training. ACM Conference'04,Month 1-2,2004.
[3] RSA online fraud report,2011. www.rsa.com
[4] Venkata Prasad Reddy, V.Radha,Manik Jindal. Client Side Protection from Phishing Attack,IJAEST 2010, vol-3,issue-1.
[5] R.Dhamija and J.D Tygar. The battle against phishing: Dynamic Security skins. In Proceedings of the 2005 symposium on Usable Privacy and security,Newyork,NY,pages 77-88. ACM press, 2005.
[6] B.Ross, C.Jackson,N.Miyake,D.BBoneh and J.C.Mitchell. Stronger password Authentication Using Browser Extensions. In 14th Usenix Security Sumposium,2005.
[7] B.Ross, C.Jackson,N.Miyake,D.BBoneh and J.C.Mitchell. A Browser Plugin Solution to the Unique Password Problem. http://crypto.stanford.edu/PwdHash/,2005.
[8] E.Kirda and C.Kruegel. Protecting Users against Phishing Attacks. Yhe Computer Journal,2006.
[9] www.yahoo.com, http://www.aol.com/, www.bestbuy.com, www.msn.com, www.paypal,com, www.ebay.com
[10]http://windows.microsoft.com/en-US/internet-explorer/products/ie/home
[11] http://www.mozilla.org/en-US/firefox/fx/

Mohiuddin Ahmed is a final year undergraduate student in Islamic University of Technology, OIC at Computer Science & Information Technology Department. Research interest includes Human Computer Interaction, Artificial Intelligence, Data Mining and Knowledge Management.

Jonayed Kaysar is also a final year undergraduate student in Islamic University of Technology, OIC at Computer Science & Information Technology Department. Research interest includes Web Engineering, Wireless Network, Human Computer Interaction and Peer to Peer overlay networks.

Study of Image Processing, Enhancement and Restoration

Bhausaheb Shinde, Dnyandeo Mhaske, A.R. Dani

Computer Science Department, R.B.N.B. College, Shrirampur Affiliated by Pune University
Maharashtra, India

Principal, R.B.N.B. College, Shrirampur Affiliated by Pune University
Maharashtra, India

Head, International Institute of Information Technology,Hinjwadi,Pune
Maharashtra, India

Abstract

Digital image processing is a means by which the valuable information in observed raw image data can be revealed. A web-based image processing pipeline was created under the ambitious educational program Venus Transit 2004 (VT-2004). The active participants in the VT-2004 can apply the basic processing methods to the images obtained by their amateur telescopes and/or they can process an image observed at any observatory involved in the project. The processed result image is displayed immediately on the display. Above that all participants can follow the distance Sun-Venus centers computation performed at the professional observatory in the real time. There is a possibility to submit an image from their own observation into the database. It will be used for the distance Earth-Sun computation.

Keywords: Educational project, WEB pipeline, image processing

1. Introduction

An image is digitized to convert it to a form which can be stored in a computer's memory or on some form of storage media such as a hard disk or CD-ROM. This digitization procedure can be done by a scanner, or by a video camera connected to a frame grabber board in a computer. Once the image has been digitized, it can be operated upon by various image processing operations**.**

Image processing operations can be roughly divided into three major categories, Image Compression, Image Enhancement and Restoration, and Measurement Extraction. Image compression is familiar to most people. It involves reducing the amount of memory needed to store a digital image.

Image defects which could be caused by the digitization process or by faults in the imaging set-up (for example, bad lighting) can be corrected using Image Enhancement techniques. Once the image is in good condition, the Measurement Extraction operations can be used to obtain useful information from the image.
Some examples of Image Enhancement and Measurement Extraction are given below. The examples shown all operate on 256 grey-scale images. This means that each pixel in the image is stored as a number between 0 to 255, where 0 represents a black pixel, 255 represents a white pixel and values in-between represent shades of grey. These operations can be extended to operate on color images.
The examples below represent only a few of the many techniques available for operating on images. Details about the inner workings of the operations have not been given, but some references to books containing this information are given at the end for the interested reader.

2. Image Enhancement and Restoration:

The image at the top left of Figure 1 has a corrugated effect due to a fault in the acquisition process. This can be removed by doing a 2-dimensional Fast-Fourier Transform on the image (top right of Figure 1), removing the bright spots (bottom left of Figure 1), and finally doing an inverse Fast Fourier Transform to return to the original image without the corrugated background Bottom right of figure 1.

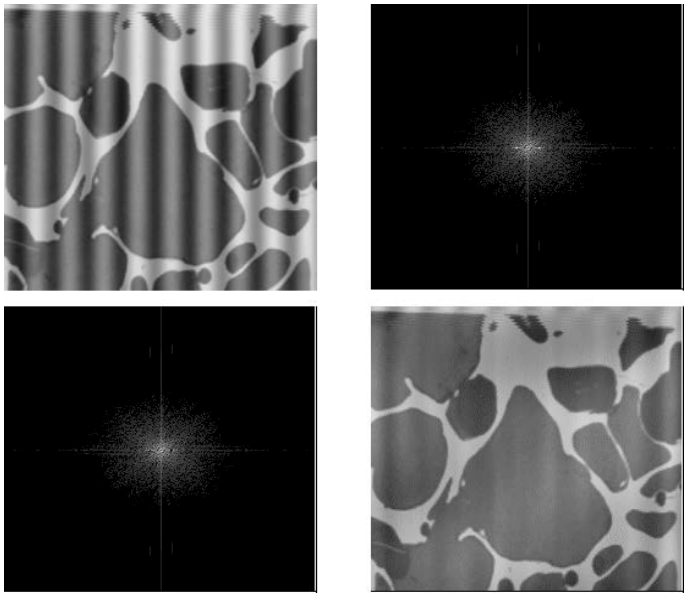


Figure 1. Application of the 2-dimensional Fast Fourier Transform

An image which has been captured in poor lighting conditions, and shows a continuous change in the background brightness across the image (top left of Figure 2) can be corrected using the following procedure. First remove the foreground objects by applying a 25 by 25 grayscale dilation operation (top right of Figure 2). Then subtract the original image from the background image (bottom left of Figure 2). Finally invert the colors and improve the contrast by adjusting the image histogram (bottom right of Figure 2.

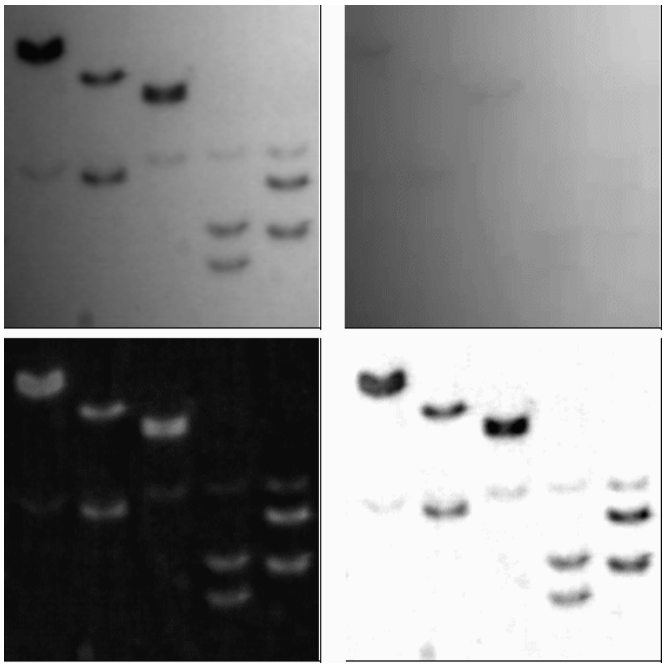


Figure 2. Correcting for a background gradient

3. Image Measurement Extraction: The example below demonstrates how one could go about extracting measurements from an image. The image at the top left of Figure 3 shows some objects. The aim is to extract information about the distribution of the sizes (visible areas) of the objects. The first step involves segmenting the image to separate the objects of interest from the background. This usually involves thresholding the image, which is done by setting the values of pixels above a certain threshold value to white, and all the others to black (top right of Figure 3). Because the objects touch, thresholding at a level which includes the full surface of all the objects does not show separate objects. This problem is solved by performing a watershed separation on the image (lower left of Figure 3). The image at the lower right of Figure 3 shows the result of performing a logical AND of the two images at the left of Figure 3. This shows the effect that the watershed separation has on touching objects in the original image.

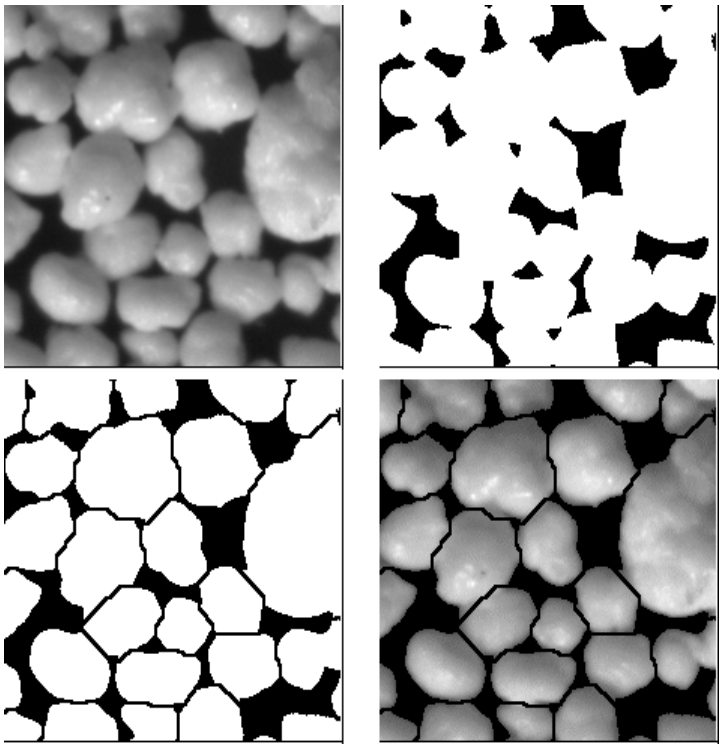


Figure 3. Thresholding an image and applying a Watershed Separation Filter

4. Color Balancing Method for Cameras:

The problem of separating the illumination from the reactance information in a given image has been extensively researched in the last three decades, following Edwin Land's seminal work on color vision and his development of the Retinex theory [4]. The problem can be described as follows – given an input image S, we would like to decompose it into two die rent images – the reactance image R and the illumination image L, such that S (x; y) = R (x; y) L (x; y). There are many benefits to such a decomposition, including the ability to correct for color-shifts due to illumination, correct for uneven illumination, introduce artificial lighting and enhancing dynamic range. It is not hard to see that in general, this problem is ill-posed – for a given input image L, there are infinitely possible solutions of L and R pairs that can explain S. Many works have tried to constraint the problem, by posing assumptions on the type of illumination (e.g. constant-hue illumination over the field-of-view and spatial smoothness). With the growing popularity of digital cameras the importance of fast algorithms for color correction (also known as auto white-balancing, AWB in short) grew as well. Such algorithms are an integral part of the image signal processing (ISP) pipeline that is responsible for converting the RAW image captured by the sensor into the final color JPEG image that is saved on the memory card. AWB algorithms try to estimate the correct three white balance gains (for the red, green and blue channels) that should be applied on an input image in order to correct for color shifts caused by illumination, so that white elements in the scene indeed appear white in the image – similar to the way the human visual system can compensate for die rent

lighting conditions so that white color always seems white under die rent illuminations. Figure 1 shows an example of correct vs. incorrect white balancing.

Conclusion: You have seen a few of the features of a good introductory image processing program. There are many more complex modifications you can make to the images. For example, you can apply a variety of filters to the image. The filters use mathematical algorithms to modify the image. Some filters are easy to use, while others require a great deal of technical knowledge. The software also will calculate the ra, dec, and magnitude of all objects in the field if you have a star catalog such as the Hubble Guide Star Catalog (although this feature requires the purchase of an additional CD-ROM).

The standard tricolor images produced by the SDSS are very good images. If you are looking for something specific, you can frequently make a picture that brings out other details. The "best" picture is a very relative term. A picture that is processed to show faint asteroids may be useless to study the bright core of a galaxy in the same field.

1.1 References:

1. Russ, John C., **The Image Processing Handbook**, 3rd ed., CRC Press
2. Jähne, Bernd, **Digital Image Processing: Concepts, Algorithms, and Scientific Applications**, 3rd ed., Springer-Verlag
3. Pratt, William K, **Digital Image Processing**, 3rd ed., Wiley
4. Gonzalez, Rafel C., Woods, Richard C., **Digital Image Processing**, Addison-Wesley

Shinde Bhausaheb: I have completed my M.C.S.(Master Of Computer Science) from Pune University, M.Phil. Also Register to Ph.D in Sighania University, Rajastan. Currently I am working in R.B.N.B. College as Head of Computer Science Department having 12 years of expert as well as Lecturer experience.

Why banks and financial institutions in Pakistan are turning towards Internet banking?

Sajjad Nazir
School of Business Management, Institute of Engineering and Fertilizer Research Faisalabad, Pakistan

Muhammad Naseer Akhtar
School of Management Studies The University of Faisalabad, Pakistan

Muhammad Zohaib Irshad
Department of Business Administration, GC University. Faisalabad, Pakistan

ABSTRACT

Internet Banking has become widespread in most developed countries, while the Financial Services Sectors in most developing countries are lagging behind with this technology. Despite the benefits afforded by such online activity, Pakistani financial institutions, in particular, have not yet experienced the full potential of this form of electronic commerce, due in part to the weakness and instability of the country's financial system. This is coupled with the fact that the citizens have lost confidence in the Pakistan Financial Services Sector in 1990s.

The objectives of this research are two-fold. The first aim is to investigate the feasibility of adopting Internet Banking within the Pakistan Financial Services Sector. The second objective is to demonstrate how Internet Banking may serve as a dual solution in restoring the viability of the Pakistan financial institutions and restoring investor confidence. From the literature review and surveys undertaken, the research examines the various benefits, which Internet Banking offers as well as its drawbacks. A comparative study reveals few reasons why financial institutions in most developing countries might not be able to embark on Internet Banking; whilst their counterparts in most developed countries are able to capitalize fully on such e-commerce venture.

The paper employs survey data to measure the extent to which financial institutions in Pakistan use e-commerce and to investigate the opportunities for further growth (that is, the likelihood of Internet Banking) within the overall Financial Sector. The research highlights a number of obstacles that must be overcome if the Pakistan financial institutions decide to actively use the Internet to provide banking services. Possible solutions that may be inaugurated to overcome the respective barriers are proposed. Finally, a summary and conclusion with recommendations are presented.

Key words: Internet Banking, E-Banking, Developing countries, Pakistan

1. Introduction:

It is widely acknowledged that Internet has permeated all types of commercial transaction in our contemporary world. The area of banking is no exception. Although the provision of banking services via the Internet is popular among developed countries (Cunningham and Froschl 1999; Jasimuddin, 2001), there exists a favourable environment for rapid development of Internet Banking to take place among developing countries as well. Internet Banking is seen to offer far- reaching potentials (Bauer, 1999), not only to the financial institutions but also to their clients and the wider society. It can enhance the institutions 'strategic initiatives and simultaneously empower customers, by enabling them to monitor their accounts 24-hours-a-day, seven-days-a-week, through the borderless environment.

Currently, it is evident that most of the financial institutions in Pakistan are employing e-commerce technologies on a wide-scale basis. They provide a combination of Automated Banking/Teller Machine (ABM/ATM) facilities, automatic funds transfer, electronic bill payment and call centre services, and with telephone banking being the latest e-commerce trend. Most of the institutions also have built websites to keep customers informed about their existing financial products and services as well as new ones that are being offered. In some cases the applications of e-commerce technologies goes beyond merely creating a presence on the Web. At least 28 commercial banks and 3 another financial institutions are now offering financial transactions over the Internet. This responsiveness to

technological innovation may prove to be a prudent course of action, considering the fragility of the Pakistan Financial Sector over the Past decade.

In the early 1990s, Pakistan began to experience a breakdown in its financial system chiefly within the Insurance and Banking sectors. By the late 1996, some of the indigenous financial institutions had collapsed. 46 of the Banks that now exist in Pakistan, almost 30 are indigenous banks; the other 16 banks are the foreign-banks. These facts, along with intermittent rumours for further closure of some of the financial institutions have raised valid concerns among the Pakistani Citizens, causing them to lose confidence in our own home-based banks and the local Financial Services Sector in general. But after 1999 the Pakistan Financial sector starts booming and the lot of foreign financial institutions started investment in banking sector.

2. REVIEW OF THE RELEVANT LITERATURE

2.1 E-COMMERCE: The Concept of Internet Banking.

Electronic Commerce (E-Commerce) is *"the application of information technology to facilitate the buying and selling of products, services and information over public standards based net works"* (Price Waterhouse Coopers (PWC), 1999). Put another way, e-commerce enables the execution of transactions between two or more parties using interconnected network. These interconnected networks can be a combination of telephone systems, Cable TV, leased lines, or wireless. E-Commerce also includes consumers making electronic payments and funds transfers (Kalakota and Robinson, 2000).

According to Howcroft (2001) "many would claims that e-commerce is reshaping almost all industries" (p. 195). This claim is true to a great extent, especially with the advent of the Internet, which has set in motion an electronic revolution in the global banking sector since 1995 (Jasimuddin, 2001). According to Bauer (1999), the financial services industry in general, and retail banking institutions in particular, were amongst the first business that realised that tremendous opportunities of the Internet and started to offer (information) services on the World Wide Web. In a similar vein, Cunningham and Froschl (1999) claim that banks are among the most intense users of technology and that the retail financial service industry deserves a special place in any discussion of electronic business.

In 1979, Hosemann predicted that by the year 2000, electronic delivery of banking services would be as commonplace as the paper cheque was in that period. Hosemann's (1979) words seem to have now come to pass. Today, Internet Banking, a form of Online Banking (Sherrod, 2000), has become a major distribution channel of banking products and services in developed world (Jasimuddin, 2001). Many Europeans banks as well as banks in the United State have been quick to embrace Electronic Business as a competitive weapon (Cunningham and Froschl, 1999). What banks are attempting to do by going online is primarily to retrain customers by reaching them more efficiently, and to increase market share (Fallenstein and Wood, 2000; Bauer, 1999; Cunningham and Froschl, 1999; Hosemann, 1979).

Online banking is broad sector that covers checking/savings/deposits, balance information, fund transfer, bill payments and credit card services (Banks, 2001). According to Banks (2001), while online banking does not possesses the 'glamour' and excitement of online trading, it is a business function that lends itself to the tools and technologies of the Internet. Internet Banking particularly allows a customer to take care of business- perform various banking tasks – using any computer that has an Internet connection and a high- speed browser (Sherrod, 2000). Another type of online banking, known as personal Computer (PC). Banking is aimed to retail customers (Banks, 2001), allowing them to use personal financial-management software, such as Quicken or Microsoft Mone, to bank from their personal computers (Fellenstein and Wood, 2000; Sherrod, 2000).

Sindell (2000) briefly points out three ways in which consumers can access their personal banking data, namely: bank-owned software using a direct dial-up, Internet access, and personal finance software. However, PC Banking remained a rather limited and cumbersome process until the commercial introduction of the Internet (Banks, 2000). Young, et al (1999) advice that a novice can go through one of the test-drive programmes which participating banks provide on their web site, in order to understand how Internet Banking works.

Figure 1

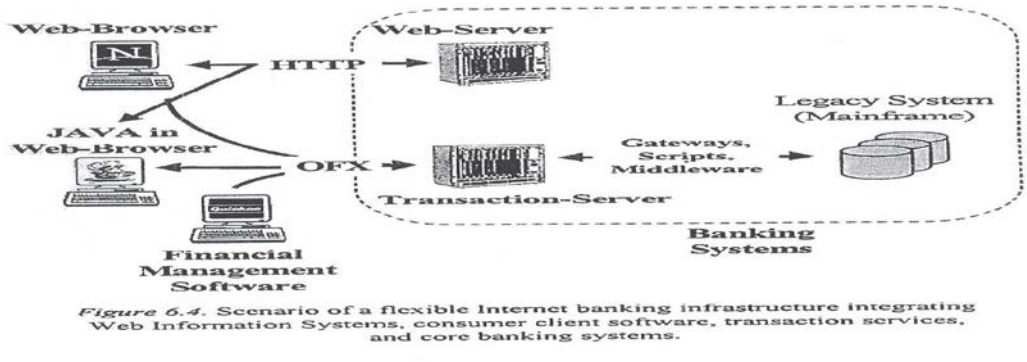


Figure 6.4. Scenario of a flexible Internet banking infrastructure integrating Web Information Systems, consumer client software, transaction services, and core banking systems.

Network Connection between Clients and Financial Institutions (Bauer, 1999; p. 72)

The benefits and the opportunities of Internet Banking coincide with the ones that e-commerce in general and the Internet in particular, bring. These benefits and opportunities are discussed in the following sections.

2.2 Benefits of Internet Banking:

The benefits of the Internet Banking are borne in the convenience it offers coupled with its enabling features. These benefits are discussed from the standpoint of the financial institutions and the consumers.

2.3 Benefits to the Financial Institutions

From an institutional perspective, many authors concur that Internet Banking is an alternative delivery channel. Strategically, banks will be continually challenged for distribution to retain their customers and market share (Fellenstein and Wood, 2000). To this end, Internet Banking offers a viable delivery channel, allowing banks to retain current customers and attract new ones, while providing improved customer service and convenience, without increasing operating costs (Humphrey's, 2000).

In addition to offering an alternative delivery channel, Internet Banking offers economic efficiencies to the institution, including "*low –cost customer service alternatives to expensive retail bank branches and telephone call centres"* (Kalakota and Robinson, 2000; p. 629). According to Humphrey's (2000), savings are realised by the bank when customers use an Internet branch to access account information and to open new accounts, minimising reliance upon personal bankers and customer service representatives for the most basic transactions.

It has been argued that the Internet presents the opportunity to level the playing field for banks of all sizes (Humphrey's, 2000), as it represents relatively low entry costs in terms of both skills and money to markets, information, contacts and culture (Miller and Slater, 2000). Therefore deregulation of financial markets lowers traditionally high entry barriers for new competitors (Bauer, 1999).

2.4 Benefits to the consumers:

From a client perspective, Internet Banking offers convenience, flexibility and significant time saving. "*Customers know that convenience and transaction simplicity count and that time is money" (Heard, 1993; p. 23).* Financial institutions recognise that customers are looking for easier ways to access information and conduct transactions; as such they see the Internet as a major commercial opportunity (Cunningham and Froschl, 1999). Internet Banking allows quick delivery of products and services. In addition, it provides control and empowerment to customers. According to Young, et al (1999), many bank allow customers to control their bank accounts online; customers can get up-to-the-minute balances on all their accounts, transfer funds from one account to another, pay recurring bills (like mortgages) automatically, or schedule transfers or payment ahead of time.

With Online Banking, commuting is reduced thus significantly saving the client's time. According to Sindell (2000) Internet Banking facilities home banking, in that it provides customers who have computer, modern, and appropriate software with the ability to download their personal bank data and conduct online activities. It is implied here that this eliminates the need for a customers to go physically to the bank and wait in a queue to conduct financial transactions. Therefore, a customer can have an

instant access to financial services and perform various banking tasks from the convenience of his computer (Sherrod, 2000). This empowerment can further lead to increased customer trust and confidence. According to Banks (2001), as customers gain greater confidence with the Internet and its delivery mechanisms, they will likely be willing to move more of their personal financial transactions to a web setting. Bauer's (1999)

2.5 Drawbacks of Internet Banking:

Doing business online has received attention for its potential, as well as for its shortcomings (Kalakota and Robbins, 2000). While the Internet may present the opportunity to level the playing fields for banks of all sizes as well as for other non-bank competitors (Humphreys, 2000), the internet at some points, will start impacting negatively on the profit margins in retail banking. According to banks (2001), the finance industry has historically been protected by high barriers to entry. However, ad financial services gravitate to the Internet some of the barberries to entry have already been eroded, thereby reducing the institutions' market share. Fellenstein and Wood (2000) point out that the biggest threat to banks from a competitive perspective is coming from the non-banking business community, typically online brokerages and software companies (e.g. Microsoft, Intuit).

On the matter of security, Baker (2000) points out that there remains fear in many places around the world about the security of the Internet transactions Hoffman (1994) explains that because the Internet is so decentralised, each computer is responsible for its own security. Therefore, there is no real inter-computer security on the Internet. As such, "*this makes it very easy for someone on the Internet to spy on transmissions undetected"* (p. 12). Assuring the privacy and preventing 'digital' fraud is a prerequisite for Internet Banking and is one of the most important factors for customer acceptance (Bauer, 1999). Huff et al (2000) use the case of Advance Bank (the first direct bank in Germany) to depict how security measures may be executed during an Internet Banking transaction. Advance Bank relies on a complex security system, where in:

"*Upon opening an account with the bank, the customer receives a Personal Identification Number (PIN) and the computer generated six-digit secret code. Every time the customer accesses his/her account by telephone or Internet, he/she is first requested to provide his/her PIN; then the bank's computer system randomly ask for 3 number from the customer's secret six digit code (e.g. the first, fourth and the fifth digits)"*(p.43).

2.6 Why Internet Banking in Pakistan?

Based on the evidence revealed in the preceding literature review, one of the most consistent arguments advanced in several Internet Banking literature is that Internet Banking is seen as a key route in increasing the financial institution's market shares and to retain their customers (see Banks, 2001); Fellestein and Wood (2000); Humphreys (2000); Bauer (1999); Cunningham and Froschl (1999). This medium of product delivery provides convenience, ease of use, low-cost transactions, and the detailed account information to both the institutions and its customers. We have seen that Internet Banking offers overall empowerment to the customer, which in turn may lead to increased customer trust, confidence and satisfaction, thereby reducing attrition. Internet Banking should enable the banks to not only provide improved service to existing clients but also to attract new customers whilst operating at a low cost. As a small, open economy, e-commerce can enable the Pakistan Financial Institutions to capture niche markets and compete effectively with larger and more developed economies to take full advantage of globalisation and free trade (PCOP, 2009).

Further justification comes from Richards (2000) who claims that when engaging in Internet business, it is an important that the business must offer products that people already desire, and that these products need to be easily transferable to and accessible on the Internet. One of the questions that must be asked is "*Does the world need your product or service?"* (p.22). According to Brigham and Gapenski (1997), commercial banks are the traditional "department store of finance" which serve a wide variety of savers and those with needs for funds (p.93).

One important fact is that the Pakistan Citizens and the world at large desire the Pakistan commercial banks, which comprises the largest share of the Financial Services Sector, offers both retail and commercial products and services that. Also, these products and services can be easily transferred to the Internet. In addition, even in its injured region, the overall Financial Sector plays a critical role in the Pakistan economy, accounting for approximately 15 - 25% of GDP (PCOP; 2010) and providing employment. Altogether, these elements may render the Pakistan Financial Sector a prime candidate for conducting its services via the Internet.

With Richards' (2000) affirmation, coupled with arguments advanced in the literature review. Internet Banking is therefore proposed as a means of strengthening the Pakistan Financial Services Sector and restoring investor confidence. However, it is carefully noted that, *"even if a company has a clear e-commerce strategy, it is not guaranteed to succeed"* (Walsham, 2001; p. 167) on the other hand, while a company is not guaranteed to succeed even if it has a clear e-commerce strategy, if the Pakistan banks choose not to face the challenges associated with these technological opportunities, *"they risk losing customers and business to faster competitors"* (Bauer, 1999; p.65).

3. Research Methodology.

The population of 106 branches of different financial institutions within the Pakistan Financial Services Sector has been identified based on information provided by the State Bank of Pakistan (SBP, 2010). The Internet Banking Survey took the form of a questionnaire along with semi-structured interviews, which consisted of a list of preset questions. This method was used for capturing in written form, a considerable amount of data from the financial institutions over a relatively short period of time. Following Haralambos and Holborn's (1990) example of administering questionnaires, the researcher gave the same questions in the same order to the respondents so that the same information can be collected from every member of the sample. The rationale here was to offer each subject approximately the same stimulus so that responses to the questions, ideally, will be comparable (Babbie, 1995 - see Berg, 2001; p. 68).

The Sector is comprised of five categorical institutions, namely: Commercial Banks, Investment Institutions, Micro finance, Islamic banks, and Development Banks.

4. Internet Banking Survey and Findings

The sample of 106 branches of different financial institutions was selected to be the subjects to whom the survey questionnaire was sent. This figure was arrived at after the author initiated telephone calls to several institutions within the Financial Sector to introduce the Internet Banking survey and to encourage their participation. However, some institutions had declined to participate due to various reasons. For instance, one of the Financial Institutions advised the author that the institution is in its preliminary planning stage for Internet Banking and does not wish to divulge any information at this tentative stage. Another instance was with the development banks. All but one advised the author that due to the nature of their business, Internet Banking is not in their future plans; hence it would make no sense for them to participate in the survey. Other financial institutions forwarded various reasons for non-participation. Nonetheless, after receiving positive responses from the other institutions, the author discovered that those institutions that wished to participate in the survey represented diverse categories of the Financial Sector, thereby providing sufficient coverage. They all ranged widely in size.

Table 1

Electronic Means of Product/Service Delivery by the Financial Institutions

Electronic Medium	Commercial Banks Total 46	Other Categorical Financial Institutions
Internet Banking	28	3
Telephone/Mobile Banking	16	2
Other*	35	2

***ATM/Debit/Credit Card/Business Card Facilities**

The above illustrates that more than 63% (28) of the commercial banks now offers Internet Banking services. Only 16 of the banks offer Telephone Banking services. 35 of the banks offer services via "Other" electronic means. "Other" primarily includes ATM/Debit, Credit card and Business card facilities. The above table also illustrates that 3-4 of the other categorical financial institutions offer product and services via the Internet, 2 Telephone/Mobile banking and 2 "Other" electronic means.

Figure: 2

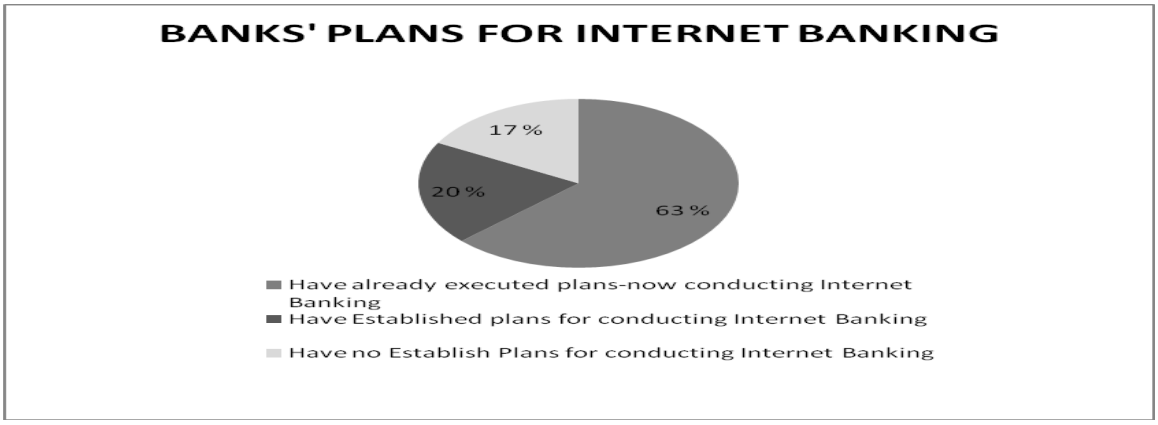


Table: 2

The Financial Institutions' Available Technological Infrastructures, Employee Awareness and Customers' Interest in Internet Banking.

Questions asked of the Financial Institutions (Commercial Banks & Other Categorical Institutions)	Yes	%	No	%
Are there available technological infrastructures?	37	80	9	20
Have employees been informed about the potential internet Banking Venture?	19	40	27	60
Has a customer/client survey been done to solicit customers' personal views on Internet Banking?	31	66.66	15	33.33

Eighty percent (80%) of the financial institutions have the basic technological infrastructure in place. Forty percent (40%) of the institutions have informed employees about the potential Internet Banking venture. Sixty-seven percent (67%) have already conducted customer/client survey to solicit customers' personal views on Internet Banking. Those financial institutions, which have conducted their customer/client survey, stated that the general consensus of the findings is that customers view Internet Banking as a good facility and as a vital banking channel. One of the commercial banks **MCB (Muslim Commercial Bank)** that is currently engaged in Internet Banking activities has pointed out that it received positive responses on all client surveys. The bank boasts that its electronic banking products are far superior in the local and even in global markets. Another bank has pointed out that its customer/client survey revealed that a high percentage of customers within the business/corporate sector require this service.

4.1 How will Internet Banking restore customer confidence in Pakistan?

In addressing the above question, the researcher first tried to identify the benefits of Internet Banking to clients and customers in Pakistan, and then tried to examine how these benefits may restore customer confidence. The survey findings show that the benefits listed by the respondents in response to the open-ended question `How *will your clients/customers benefit from Internet Banking services?* Concur with those identified in the literature review.

Table: 3

Benefits of Internet Banking to Clients/Customers

Convenience	• Single access point for all financial products and services. • Banking at customers' own convenience.

	• Ensures better monitoring. • Portability.
Accessibility (Easy Access)	• Global access to accounts; clients can access account information anywhere and at anytime (24-hours-a-day; seven-days-a-week). • Higher availability of bank data.
Increase Competition	• Give local merchants a chance to compete on international markets.
Increase Profitability and Savings	• Merchants/corporate clients get funds of varying currencies. Deposited to their local accounts. • Ordinary citizens can reap similar benefits.
Saves Time	• Less time required for bank business. • Quick delivery of products and services. • Reduces commute.
More Choices	• Can select from many financial institutions and from more products and Services
Possibility of Cost-savings	• Using “cheaper" delivery channels.

From the above, we can construe that the benefits of Internet Banking in themselves are ideal factors for restoring and maintaining investor trust and confidence in Pakistan.

An evidence from MCB Bank that people tend to use technology, as it is convenient to them. They use this facility given that they don’t have time to address personal affairs during weekdays - by the time they get home from work the banks are closed. In addition to this, while putting credit card details over the Internet may involve a few risks, the transaction convenience far outweighs these risks.

Taken altogether, it can be construed that Internet Banking promotes quick response, convenience, and improved quality which give rise to other benefits such as time saving, cost-savings, easy access, wider choices and customer empowerment. All these benefits in turn are geared toward satisfying the customers and clients. Customer satisfaction invokes feelings of gratification, thereby replacing fear and mistrust with confidence and trust. According to Gibson et al (1997), a satisfied customer will continue to repeat business with a particular organization. Therefore, if the financial institution’ products and services adequately meet customers’ needs then this can restore, boost and maintain their confidence in doing transactions with the respective institutions. In addition, the current reliance on technology within the Pakistan society indicates that there will be some amount of commitment by investors and the citizens in general, to participate in Internet Banking.

4.2 How will Internet Banking strengthen the Pakistan Banking Industry and the Local Financial Services Sector in general?

Another observation made from *table*, which is in accordance with Richards’ (2000) claim, is that most of the products and services offered by the Pakistan financial institutions could be easily transferred to the Internet. For example, customers may apply for loans, credit cards and may make loan, credit card repayments online. Some of the potential benefits of Internet Banking to the financial institutions as stated on the returned questionnaires are summarized in Table 4.

Table: 4
Benefits of Internet Banking to the Financial Institutions

Benefits of Internet Banking to the Financial Institutions
➢ Increased relationship with customers, giving rise to greater loyalty and share of wallet
➢ More cost-effective mechanism for communicating with customers
➢ Improved banking services
➢ Less staff (e.g. tellers and customer service officers) and less office space
➢ Able to reach a wider cross-section of customers. Reach more offshore customers as they would be able to view account information from anywhere in the world.
➢ Provides revenue and increases profitability
➢ Provides real-time banking information to customers
➢ Reduces the need for branch expansion, more reach and availability without an investment in property
➢ Decreases downtime if access workstation is affected

If the Pakistan financial institutions are to conduct their services via the Internet then this could mean immediate expansion of their marketplace to national and international markets. It will also provide them with the opportunity to reach their customers more efficiently. As mentioned earlier, quick response and improved quality promote customer satisfaction. Customer satisfaction, as affirmed by Gibson et al (1997), is the key to organizational success *"for it is the satisfied customer who accounts for the repeat business that the organizations need to survive and thrive" (p. 214).* Therefore the Internet as a medium of product delivery would enable the institutions to not only provides improved services to existing clients and customers, but also to retain them and attract new ones whilst operating at a low cost. This, along with customer reliance upon the institutions' products and services may help to restore the viability of the financial institutions and strengthen their business, as success in e-commerce will have an immediate impact on the institutions' productivity and profits.

4.3 The Future of Internet Banking in Pakistan

One of the main conclusions that can be drawn from the survey is that Internet Banking in Pakistan is highly feasible. It has been revealed that financial institutions (commercial banks and other financial institutions from diverse categories) have already begun the Internet Banking venture, whilst the others seem to be making extensive preparation for this type of e-commerce business. This responsiveness to technological innovations can enhance the financial institutions' strategic initiatives and simultaneously re-establish some amount of confidence among the Pakistan citizens by allowing clients and customers to monitor their own financial accounts 24-hours-a-day, 7-days-a-week through a borderless environment.

The author has also seen from the findings that the benefits of Internet Banking in themselves are ideal factors for restoring and maintaining customer trust and confidence in Pakistan. For example, beneficial factors such as quick response, convenience, and improved quality give rise to other benefits such as time saving, cost-savings, easy access, wider choices and customer empowerment. All these benefits in turn lead to customer satisfaction. Customer satisfaction invokes feelings of gratification, which will motivate customers to repeat business with a particular financial institution, thereby replacing fear and mistrust with confidence and trust. Customer satisfaction is the key to organizational success, as it is the satisfied customer who accounts for the repeat business that the organizations need to survive and thrive. We may therefore conclude that Internet Banking can serve as a dual solution in restoring the viability of the Pakistan Banking Industry and restoring customer confidence.

Other factors, which support this conclusion, are the emphasis that Pakistani's are now placing on technology and the citizens' desire for maintaining financial products and services. It can be speculated that as the Pakistan financial institutions continue to

upgrade and refine their e-commerce strategies, this will encourage and provide opportunities for other institutions within the Financial Sector to embark on these e-commerce ventures as well. This in turn will strengthen the overall Pakistan Financial System.

5. Potential Benefits to Pakistan

Internet Banking will not only benefit the Pakistan financial institutions and their respective customers but will impact positively on the entire country as well. For example, it may help to provide new jobs, employment and livelihoods for the Pakistan citizens. The survey findings show that most of the benefits listed by the respondents in response to the open ended question '*How can Pakistan benefit from Internet Banking?* These are summed up in *Table 5 below.*

Table: 5

	Benefits of Internet Banking to Pakistan
International Reach	• Providing a new way for local entities to do business overseas and fulfilling cross-border banking needs. • Potential for more investments locally by citizens living outside of Pakistan.
Sophistication of Basic Infrastructure.	• Increased technology exposure for citizens generally adds to the sophistication of basic infrastructure of the country thereby increasing its appeal to the investment community. • Better image for Pakistan, particularly as a technology destination with superior financial services.
Increased Competition	• Gives local merchants a chance to compete on international markets as well as provide a more competitive industry to global clients.
Additional area for Resource Development	• Generation of new jobs/employment, new job skills, and livelihoods.
Increased Productivity and Reduced Pollution	• More production time as less people will need to leave work to go to the bank. • Less commuting would reduce the pollution from the motor vehicles, as there will be less traffic on the road.
E-commerce Growth	• Enabling/paving the way for further development of e-commerce and merchant commerce (m-commerce) activity.
Convenience and Possibility of Cost Savings	• Provide another (less expensive) alternative for consumers and business to conduct their business.

6. Conclusions.

Internet Banking is a very marginal activity in Pakistan, as most of the financial institutions have not yet experienced the full potential of this form of e-commerce. Only 28 commercial banks and 3 other categorical financial institutions have already embarked on the Internet Banking venture. Furthermore, this type of service is currently being offered predominantly to merchants and corporate clients (business-to-business more than business-to-consumer). The other institutions within the Pakistan Financial Sector generally use the Internet to create an electronic presence and to keep their customers informed about the institutions' existing as well as new products and services. However, the possibility of rendering transactional banking services via the Internet in Pakistan remains quite high, as the research findings revealed that most of the financial institutions are interested in this venture and have begun extensive planning.

Whilst Internet Banking is still a novelty among the Pakistan mass, most the consumers have been experiencing the benefits of telephone banking services. These benefits are similar to those derived from Internet Banking. For example, access to accounts information, bill payment, and fund transfer are available 24 hours and can be done at the customers' convenience from anywhere in the world, using a standard telephone.

Internet Banking as a possible e-commerce solution will create possibilities for local financial institutions by marketing their products and services, thereby attracting new clients and customers. The prescribed databases will link to international information sources, and online

information will be provided to the institutions' clients and customers. This not only may meet the needs of the customers but also restore confidence and improve the quality of life for all citizens.

While Internet Banking is not without its drawbacks and challenges, this research recommends that the Pakistan Financial Sector should move contiguously towards an e-commerce solution by conducting its transactional banking services through the Internet.

References

Baker, Gill (2000) *Tech Obstacles Bar China's Road to Web Commerce: The People's Republic is a Long Way from Jumping on the e-financial Bandwagon. Global Tec watch - Bank Technology News* 14(7), pp. 52.

Banks, Erik (2001) *E-Finance: The Electronic Revolution in Financial Services.* **John Wiley & Sons, Ltd.**

Bauer, Christian (1999) *Financial Institutions and the Internet: Issues and Trends.In:* **F. Sudweeks and C. Romm (eds.)** *Doing Business on the Internet: Opportunities and Pitfalls.* **Springer-Verlag London Ltd. p. 65-75**.

Brigham, Eugene and Gapenski, Louis (1997) *Financial Management: Theory and Practice.* **8th edition**. **The Dryden Press.**

Citibank of Bank of Pakistan. (Online) Available from http://www.citibank.com/pk/products services/commercial banking

Cunningham, P and Froschl, F (1999) *Electronic Business Revolution: Opportunities and Challenges in the 21st Century.* **Springer-Verlag Berlin Heidelberg.**

Fellenstein, Craig and Wood, Ron (2000) *Exploring E-commerce, Global E-business, and E--societies.* **Prentice Hall Inc.**

Gattiker, Urs E. (2001) *The Internet as a Diverse Community: Cultural, Organizational and Political Issues.* **Lawrence Erlbaum Associates, Inc., Publishers.**

Gibson, J., Ivancevich, J., Donnelly, J. (1997) *Organizations: Behavior, Structure, and Processes.* **9th edition. Times Mirror Higher Education Group, Inc., Company.**

Government of Pakistan (2010) *A Five-year Strategic Information Technology Plan for Pakistan*.

Heard, Ed (1993) *Walking the Talk of Customer Value.* **National Productivity Review/Winter, 1993/94.**

Hoffman, Paul E. (1994) *Internet: Instant Reference.* **SYBEX Inc.**

Hosemann, Michael J. (1979) *"The Rationale for Electronic Banking"* **American Bankers Association.**

Howcroft, Debra (2001) *After the Goldrush: Deconstructing the Myths of the Dot.com Market.* ***Journal* of *Information Technology*, 16, pp. 195-204.**

Humphreys, Kim (2000) *Internet Banking: Leveling the Playing Field for Community Banks.* **In: J. Keyes (ed.)** ***Financial Services Information Systems.*** **p. 621-629.**

Jasimuddin, Sajjad M. (2001) *Saudi Arabian Banks on the Web. Journal of Internet Banking and Commerce, 6(l).* **Available from:** http://www.arrgydev.com/commerce/jibc/0103_02.htm

Kalakota, Ravi and Robinson, Marcia (2000) *Electronic Commerce. Encyclopedia of Computer Science.* **4th edition. Edited by Ralston et al., Nature Publishing Group.**

State Bank of Pakistan (2010). Online available from (www.sbp.org.pk). Last accessed 10 December 2010.

An Authoring System for Editing Lessons in Phonetic English in SMIL3.0

Merzougui Ghalia[1] and Djoudi Mahieddine[2]

[1] Departement of Informatics, University of Batna, 05000, Algeria

[2] Laboratory XLIM-SIC and IRMA a Research Group, UFR Sciences SP2MI, University of Poitiers Teleport 2, Boulevard Marie et Pierre Curie BP 30179 86962 Futuroscope, Chasseneuil Cedex- France

Abstract

One of the difficulties of teaching English is the prosody, including the stress. French learners have difficulties to encode this information about the word because it is irrelevant for them. Therefore, they have difficulty to produce this stress when they speak that language. Studies in this area have concluded that the dual-coding approach (auditory and visual) of a phonetic phenomenon helps a lot to improve its perception and memorization for novice learners. The aim of our work is to provide English teachers with an authoring named SaCoPh for editing multimedia courses that support this approach. This course is based on a template that fits the educational aspects of phonetics, exploiting the features of version 3.0 of the standard SMIL (Synchronized Multimedia Integration Language) for the publication of this course on the web.

Keywords: *authoring, document model, mediated courses, phonetics, dual coding, paralinguistic markers, SMIL3.0.*

1. Introduction

The English and French share a large lexicon, where the spelling forms of a word in both languages are similar. However, the accentual system established by the two languages for these words make them opaque for oral learners. It is observed that certain syllables are more readily audible than others. We speak about accented syllables in this case.

The French learner faces two challenges: to perceive, in the listening phase, the accented and unaccented syllables, and reproduce during the production phase, a sufficient contrast between the two types of syllables.

French students may have serious gaps in phonetic and prosodic when making oral presentations, despite a correct language on the lexical and syntactic. The absence of discrimination vowel/diphthong and the displacement of stress make some words unrecognizable.

Controlled empirical studies confirm what teachers observe every day. These studies show that Anglophone Canadians recognize less isolated words pronounced by a French Canadian. The authors attribute this difference to a lack of emphasis. But [6] confirms that this problem is not sensory but lies at the level of working memory; there is negligence when encoding information. French students learning English fail to treat stress, which has little value in their native language, and therefore they do not store this information. During their presentations, they will put the accent on a random syllable of an English word, and this indicates negligence in encoding and a lack of storage of the place of the stress and not deafness or a production problem. This has a negative effect on understanding their speech.

There are some students who after 10 years of learning English language, have not yet mastered the pronunciation of words that seem elementary (like: who, women, chocolate, village, low, allow, sun, sound). Series such that (there're, aren're, were not, were, where) or graphically similar words such as (tough, trough, though, through, thought) pose enormous problems of memorization in oral.

In [7], Beck et al have made the hypothesis that attention processes play a fundamental role in this case. The visual computing solutions seem to be a good solution. Visual tracking helps in distinction of parts of speech in which problems of perception and understanding arise. The sound becomes visually observable in time, unlike its first material form of transient vibrations of air. The pronunciation would be easier if the student simultaneously read and hear the word 'development', where the stressed syllable is underlined visually (a different style, font and color associated). Treating such a word is in auditory encoding of the linguistic information and in a visual encoding of both language and paralinguistic information, hence the need for a tool for editing documents supporting such a presentation.

In section two, we present the tools for pronunciation and the tools for editing multimedia documents in SMIL standard, and we will position our system with respect to these two groups of tools. Section three presents the model of document proposed to support the approach of dual

coding of phonetic aspects to teach. Then, in section four we describe the architecture of our authoring. The article ends with a conclusion and some perspectives.

2. State of the Art

2.1 Computing Tools of Pronunciation

In the recent years and as regards the acoustical phonetic, advances in graphic display screens are spectacular. Software such as (speaker, tell me more, English plus, book or voice) using oscillo-grams to present the voice have a limited supply, while (Win Pitch, Speech analysis), using curves of fundamental frequencies, are too limited in use because of their complexity or because of ergonomics errors.
In [9], the authoring Sound Right was based on the basic curves for drawing simplified intonative curves using extensible arrows that appear below the text. The difficulty of interpreting complex curves explains their limited use in language courses in schools and universities.
Prosodic Font is a system developed at MIT Media Lab [12] to automatically generate dynamic fonts (from an oral speech input) that vary with the time and the change of tone in a speech. The goal is to generate animated text depending on the intonation and on the prosody of speech. Such a solution is not accepted as the best didactic solution for teaching pronunciation.

2.2 Publishers of Multimedia Course in SMIL Format

In this section we briefly describe three editors of SMIL documents:
SWANS [7] is an authoring system that allows any teacher to semi automatically generate multimedia documents where the accent is marked visually (by typographical markers such as color style ...) and aurally. The generated document is a web page where the learner can read and/or listen to a speech synchronized with the text annotations. The scenario requires initial import of media (text, audio and video) in the work environment, then, synchronization of the text (which is segmented into units of breath) with its audio or video pronunciation and finally, the teacher can annotate the text by typing markers.
The system LimSee3 [2] is a multimedia publisher of new generation, which uses templates to simplify editing and ease repetitive tasks. It also allows users to generate documents for different output formats (SMIL, XHTML + JavaScript and timesheets). Currently, there are three templates that are based on the construction of a multimedia course:
The first model allows to build a slide show (set of slides) to prepare a course. Each slide can contain one or more media. These media are inserted or imported from outside by simple gestures (copy and paste or drag and drop).
The second template enables his user to build a course record. In this case, the issue needs first to import the slides used during the lesson or the teacher's image and voice (video and audio track), then to synchronize them by adjusting the transparencies with audio and video. The synchronization tool allows replay of the audio plug and indicates by clicks the times when we need to change the transparency.
The third model can annotate in real time, oral examinations of students.
The publisher ECoMaS [8] is also an editor of mediated courses and is based on document's models. The final presentation is generated in the same way as in the second model LimSee3, but the scenario of edition is different. The editing by ECoMaS requires to import transparencies which are images, then one has to record the oral explanations of each, then the system generates a publishable presentation on the web (SMIL2.0)[11], where slides are synchronized with their audio explanation and an index table which provides temporal navigation during the presentation of the course.

The last three editors use document templates organized hierarchically. Each model is seen as a document with holes, where the user simply fills the holes by media (text, image, audio or video). It is clear that these Medias are imported from external sources and therefore the teacher must prepare in advance each of them with its corresponding tool. This is tedious, especially if it is to edit formatted text (with colors and styles ...) and then associate it with pronunciation; he must use two different tools (one for word processing and one for the sound), import them into the system and then synchronize them.
An annoying limitation of these tools is the lack of a graphics tool to edit formatted text. Versions of SMIL 1.0 and 2.0 used by these tools do not support tags for text formatting (color, style, font ...), which is very important to materialize the dual coding approach mentioned in our problem. SWANS uses standard XHTML + SMIL. ECoMaS uses the language RealText only to make the title of a transparent and the content of the index table.
The latest version of SMIL [4] supports text formatting features within document (.Smil), but so far there is no graphical editor for this version. Our work is the first contribution to the development of such an editor.
Using the International Phonetic Alphabet (IPA), phonetic attempts to represent the sounds more accurately but teachers cannot use this type of character with existing publishers.

3. The SaCoPh Approach

3.1 Objectives

-Improve the collection and the storage of phonetic concepts (such as stress) using the dual-coding approach;
-Provide language teachers with an editor that allows the preparation of multimedia course publishable on the Web (SMIL3.0 standard) as a template that fits the teaching of phonetic concepts. This tool must be flexible via a more user-friendly interface, and it must be as close as possible to the principle of WYSIWYG.

3.2 Course Models

Our proposal is that a course of phonetics must be composed with a series of lessons where each is represented by a multimedia document; so the lesson will be generated by our tool as follows:
Because each lesson is a multimedia document, its description takes into account its various dimensions:

- Structural dimension

Each lesson contains a title and a set of phonetic rules accompanied by examples. Each rule will have demonstrative examples, and each rule or example will be associated with its pronunciation (audio file). Parts of the text of an example on which the teacher wants to attract the attention of the learner are highlighted visually (wear a different color and style).

- Spatial dimension

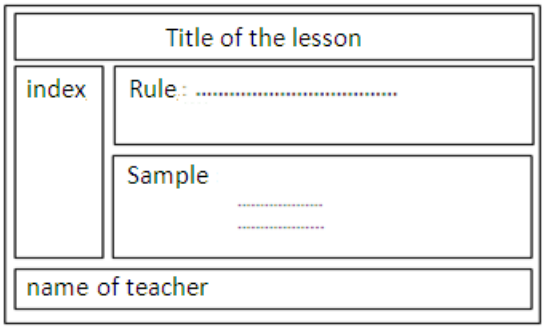


Fig. 1 Spatial dimension of a lesson.

- Temporal dimension

Each rule or example appears in parallel with its pronunciation, and they appear in sequence. The rules will follow in time one after another. The following figure shows the temporal aspect of the document.

The media objects are represented by rectangles where the length reflects the duration of display or presentation of the corresponding object.

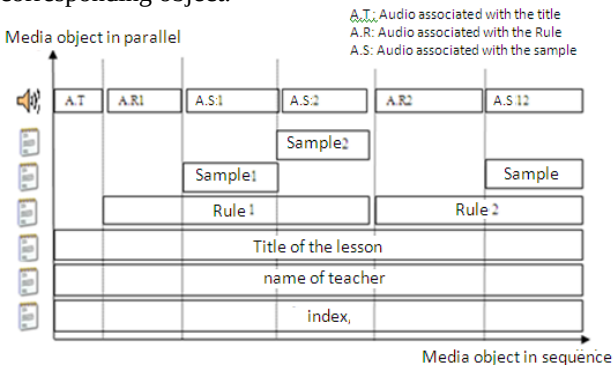


Fig. 2. Chart shows the temporal aspect of a document

- Hypermedia temporal dimension

The index object contains a list or summary of the rules of the lesson. The elements of this list are clickable areas where a click on one of them enables to watch the presentation of the lesson at the beginning of the corresponding segment (the rule in question); it represents time navigation.

4. Development of SACoPh System

The approach of our system is to associate a modality of typographical representation with an oral modality of accentuation. The combination of typographic style to the quality of spoken discourse has been little explored. There are three ways to combine these two types of representation: one qualified as automatic (the system Prosodic Font), one described as interactive and the last combining the two solutions (the system SWANS for example).
In our case, the solution is interactive because it allows the author (teacher) to choose its own submission on one hand and to decide where he wants to focus on the other.

4.1 Data Structure

The data structure of a lesson is presented by the class diagram as follows (using the UML):

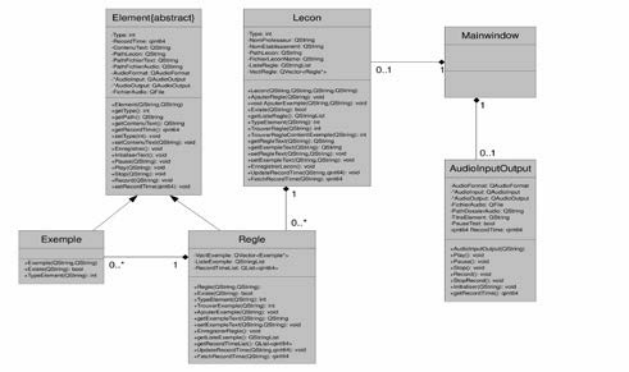


Fig. 3. Diagram of classes of a course document

4.2 System Architecture

The system SaCoPh consists of four modules presented in Figure 4.

Management of the data structure: This module offers to teachers the opportunity to manage the lesson by managing a tree. It can create, edit or delete a rule or an example. Deleting a rule therefore remove all the examples it contains.

Word Processing (typographical marker): Through the features of this module, the user can enter the text (of a rule or an example), insert the phonetic alphabet characters by selecting them from a list, and the most important, it can add typographic markers (color, form, style and/or size) on parts on which he wants to attract the attention of the learner. This module generates XHTML code that corresponds to such a specification for later use, to update or to facilitate the generation of SMIL code thereafter.

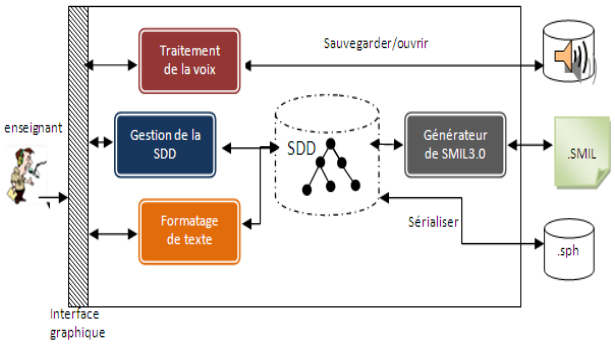


Fig. 4. SACoPh Architecture

Voice processing: This module is concerned with recording the voice of the teacher in an audio file format 'wav'. It enables to replay or to delete what was recorded and it provides information about the object as duration.

Generating of a SMIL presentation: A document published by our system is saved (serialized) into its own format (.Sph), and is ready for later updates. Nevertheless, the teacher can export documents via this module to the SMIL 3.0 format for publish or share. So far, only the player called Ambulant can read the presentations of version 3.0. Thus, this module must traverse the tree of object rules and examples (SDD) to generate the SMIL code. The synchronization between these objects is implicitly deducted from their order in the tree. The display duration of a rule or an example is deducted from the duration of the object associated.

```
<-- specify the time segment containing a rule
and its samples -->
<par xml:id="1" dur="28s">

<-- specify the pronunciation of a rule and its
samples -->

  <audio begin="1s" src="Regle 1.wav"/>
  <audio begin="11s" src="Exemple1_R1.wav"/>
  <audio begin="14s" src="Exemple2_R1.wav"/>
  …
  <-- specify a rule 1 -->
  <smilText region="Regle">
   <p>
   <span textFontFamily="…" textColor="#..."
       textFontSize="16px"> The vowel </span>
   <span textFontFamily="…" textColor="#..."
       textFontSize="18px"> a </span>
   <span textFontFamily="…" textColor="#..."
       textFontSize="16px"> is pronounced ...
       </span>
   </p>
   </smilText>

<-- specify the sample 1, witch starting in the
11th second after Rule 1 -->

   <smilText begin="11s" region="Exemple">
    <p>
    <span textFontFamily="…" textColor="#..."
        textFontSize="16px"> W </span>
    <span textFontFamily="…" textColor="#..."
        textFontSize="18px"> a </span>
    <span textFontFamily="…" textColor="#..."
        textFontSize="16px"> tch </span>

<-- specify the sample 2: witch begin at the
third second after the sample 1  -->
   <tev begin="3s"/>
   <p>
    <span textFontFamily="…" textColor="#..."
        textFontSize="16px"> B </span>
    <span textFontFamily="…" textColor="#..."
        textFontSize="18px"> a </span>
    <span textFontFamily="…" textColor="#..."
        textFontSize="16px"> th </span>
   </p>
   ...
  </smilText>
</par>
…
```

We show above some of the code generated by SMIL3.0 SACoPh to synchronize the text of a rule in

parallel with pronunciation. Note that after 11s, the text of the first example is presented with its pronunciation. Note the different types of tags and attributes: those used for synchronization <par>, <tev>, begin and those used for the presentation of typographical marker <smilText>, textFontFamily, text Color, etc.

An index table is calculated automatically. We proposed a temporal segmentation of the document. This is to identify the start and duration of each segment, then place markers for each of them to be referenced later. We propose that a time segment is the time for filing a rule with all its examples: this segment is an indivisible entity. We made this choice because the rule can not be fully understood only through its explanatory examples, and they can not be divorced from the rule. Each entry in the index table, we associate a link to its corresponding time segment through its marker. The following code shows how the segments are marked, and how are referenced by entries in the index table.

```
<-- specify the sequence of temporal segments and
mark each by the tag xml :id-->
<seq>
   <-- segment 1 -->
   <par xml:id="1" dur="28s">
     <smilText region="Regle">
       ……
     </smilText>
     <smilText region="exemple">
       …
     </smilText>
       ……
   </par>
   <-- segment 2 -->
   <par xml:id="2" dur="15s">
       ……
   </par>
   ……
</seq>
  <-- complete the index table -->
   <a href="#1">
    <smilText region="Index1"> Rule 1</smilText>
   </a>
   <a href="#2">
    <smilText region="Index2"> Rule 2</smilText>
   </a>
   ……
```

The index table provides a navigation time during the show of the lesson. Students can forward or rewind the presentation to the beginning of a rule he wants replay by clicking the link in the table. Figure 5 shows the final presentation of a lesson by the ambulant player.

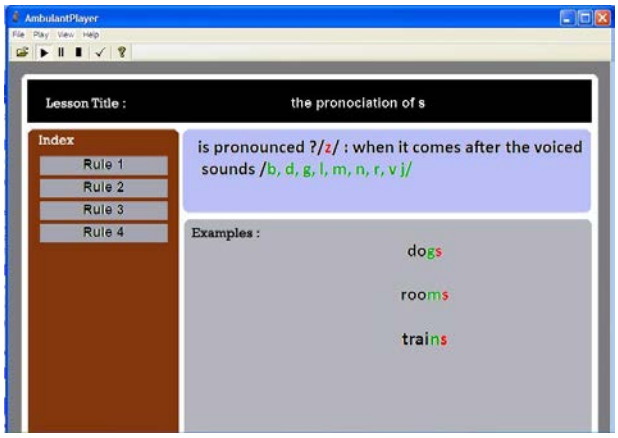


Fig. 5. – présentation d'une leçon générée avec ambulant

4.3 Interface SaCoPh

Our system is developed in C + + using the QT library; it allows a development of cross-platform graphical applications based on the following approach: write once and compile anywhere. All the services supported by our editor are provided via a graphical interface; it is easy to use and it takes into account the already entrenched attitudes among teachers. The figure below illustrates an overview of this interface.

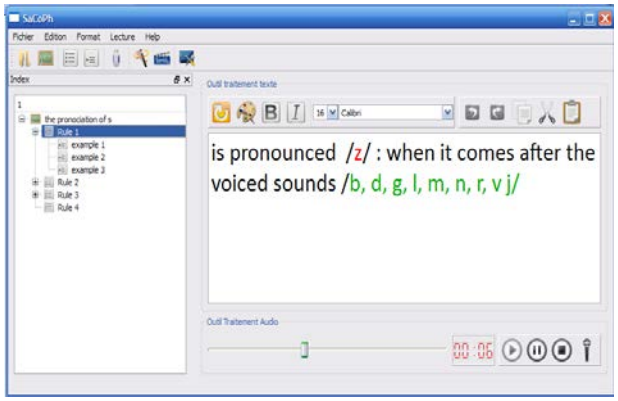


Fig. 6. GUI of SACoPh

5. Conclusions

Our research focus is based on the processing of multimedia documents and timed applied in distance education, and more specifically to the teaching of phonetics of a language. We presented the significant contribution of the dual coding approach in improving learning. For this, we designed a template that concretizes

this approach. We have developed an authoring system called SaCoPh, which generates over phonetic publishable via the Web, according to this model, using the new features of version 3.0 of the SMIL standard. This system is designed for non-computer for teachers, it provides a simpler interface and more user-friendly as possible.
Moreover, we consider the following perspectives:
Integrate into our system functionality that allows to record video and/or draw the image. That way, the teacher will not have to use various external tools to prepare different types of media when preparing his lessons. If he wants to use an existing media, it can import the URL
Accelerate the process of publishing with a semi-automatic thinner synchronization mechanism and especially when the media is imported.
Facilitate for learners, the research via segments in lessons throughout the course of this format.

Acknowledgments

G. Merzougui would like to thank a lot both M. Moumni and M. Aouadj for discussions, comments and suggestions that have greatly enriched the work.

References

[1] J. Mikač and C. Roisin and B. Le Duc," The LimSee3 Multimedia Authoring Model". ACM Symposium on Document Engineering, 10-13 October 2006, Amsterdam, The Netherlands, pp. 173-175
[2] Jan Mikác, Cécile Roisin « Comment bâtir un cours multimédia avec LimSee3 ? » EpiNet : Revue électronique de l'EPI (2008)
[3] www.ambulantplayer.org
[4] Synchronized Multimedia Integration Language (SMIL 3.0), W3C Working Draft 13 July 2007, http://www.w3.org/TR/2007/WD-SMIL3-20070713/
[5] R. Deltour and A. Guerraz and C. Roisin, « Multimedia Authoring for CoPs ». 1st International Workshop on Building Technology Enhanced Learning solutions for Communities of Practice, Crete, Greece, 2 October 2006, pp. 60-69
[6] Antony Stenton, Anne péchou, Christine Caillant-Sirdey et André Tricot «Effet du double cordage synchrone de l'accentuation en L2 selon des modalités de restitution de l'apprenant » 1er colloque international de didactique cognitive, toulouse, 26-28 janvier 2005.
[7] Aryel Beck et al, « SWANS, un système auteur de synchroisation et d'annotation pour un apprentissage multimodale de phénomène accentuels en langue vivante L2 » www.lairdil.org/publications/Swans_1_2005_N._Publie.pdf
[8] Ghalia Merzougui, Mahieddine Djoudi, Abdelmadjid Zidani, "Editeur de cours médiatisés en SMIL", Conférence Internationale: Sciences Electroniques, Technologies de l'Information et des Télécommunications, IEEE SETIT 2004, ISBN 9973-41-902-2, Sousse, Tunisie, 15-20 Mars 2004.
[9] Péchou, A. Senton «Encadrer la médiation- le cas de l'intonation » colloque compréhension et hyermédia, Albi, Octobre 2002.
[10] http://www.w3.org/TR/2002/NOTE-XHTMLplusSMIL-20020131/
[11] http://www.w3.org/TR/SMIL2/
[12] Tara Rosenberger, Ronald L. MacNeil «Prosodic Font: Translating speech into graphics» Proceedings of CHI'99 Extended Abstracts. http://www.media.mit.edu/~tara/CHI1999.pdf.

Merzougui Ghalia received a Master in Computer Science from the University of Batna, Algeria, in 2004. She is currently a Professor at the University of Batna, Algeria.
She is a member of (Adaptive Hypermedia in E-learning) research group. She is currently pursuing his doctoral thesis research on the management of multimedia educational content. Her current research interest is in E-Learning, system of information retrieval, ontology, semantic web, authoring and multimedia teaching resource. His teaching interests include computer architecture, software engineering and object-oriented programming, ontology and information retrieval.

Mahieddine Djoudi received a PhD in Computer Science from the University of Nancy, France, in 1991. He is currently an Associate Professor at the University of Poitiers, France.
He is a member of SIC (Signal, Images and Communications) Research laboratory. He is also a member of IRMA E-learning research group. His PhD thesis research was in Continuous Speech Recognition. His current research interest is in E-Learning, Mobile Learning, Computer Supported Cooperative Work and Information Literacy. His teaching interests include Programming, Data Bases, Artificial Intelligence and Information & Communication Technology. He started and is involved in many research projects which include many researchers from different Algerian universities..

TBEE: Tier Based Energy Efficient Protocol Providing Sink and Source Mobility in Wireless Sensor Networks

Siddhartha Chauhan[1] and Lalit Awasthi[2]

[1]Department of Computer Science and Engineering, National Institute of Technology, Hamirpur, H.P. 177001, India.

[2]Department of Computer Science and Engineering, National Institute of Technology, Hamirpur, H.P. 177001, India.

Abstract

In resource constrained wireless sensor networks (WSNs) it is important to utilize energy efficiently. Data dissemination is mainly responsible for the consumption of energy in sensor nodes (SNs). The data dissemination protocols for WSNs should reduce the energy consumption of the SNs. Sink and source mobility is the major challenge for data dissemination protocols. In this paper, a Tier based Energy Efficient protocol (TBEE) providing sink and source mobility in WSNs has been proposed. TBEE protocol has been designed so that fewer SNs located nearer to the dissemination point (DP) respond to the sinks message for grid formation thereby reducing message overheads. TBEE exploits an improved approach of communication amongst the SNs so that the collisions are reduced. TBEE efficiently handles the movement of the sinks and sources in the network and reduces the overheads associated with their mobility. TBEE's performance was evaluated in different conditions and scenarios. Simulation results show substantial improvement by TBEE as compared with the other existing grid based approaches for most of the scenarios.

Keywords: *Wireless sensor Networks, Grid based data dissemination, TBEE.*

1. Introduction

WSNs are being used for applications such as agriculture, habitat monitoring, military surveillance, security intelligence and industry automation. The various challenges of WSNs are scalability, fault tolerance, hardware, power consumption, and topology change [1]. These challenges are to be dealt in order to provide better and efficient solutions for different applications of WSNs. Energy optimization is very important as it improves the life time of WSNs. Reducing energy consumption for extending the lifetime of WSNs is a challenging task [3].

The main purpose of data dissemination is not only to transmit information related to data or query; but also to reduce the overall energy consumption [4, 5]. A number of protocols have been proposed to achieve reliable data dissemination in WSNs. The direct data dissemination approaches are the fastest and easiest but are only feasible for static networks; where SNs have information of the other nodes. Single hop transmission used by direct data dissemination approaches is highly impractical for WSNs. Multi-hop data dissemination protocols support the cooperative effort of various SNs for data dissemination. SNs have a transmission range referred as the distance; where the signal strength remains above the minimum available level for a particular SN to transmit and receive data [6]. If two SNs are not capable of direct communication, they route their data through the intermediate nodes between them [7].

This paper mainly focuses on the mobility (sink and source) and energy efficiency for data dissemination in WSNs. In this paper, we have proposed a protocol namely tier based energy efficient protocol (TBEE) for static SNs where sinks and sources change their locations dynamically. The performance of TBEE has been analyzed and compared with two tier data dissemination protocol (TTDD) [2] and grid based data dissemination protocol (GBDD) [22]. The effect of the grid size on TBEE in terms of energy consumption has also been analyzed.

The rest of this paper is organized as follows: In Section 2, several related work are introduced. In Section 3, we present the analytical model with equations for energy and message overhead calculations. In Section 4, we present our proposed tier based energy efficient protocol. In Section 5, performance of TBEE has been analysed and compared with existing grid based data dissemination schemes. Section 6 is of conclusion.

2. Related Work

Energy efficient routing and data dissemination are the

most highlighted research issues in WSNs these days. Sensor protocols for information via negotiation (SPIN) [8] is an important work that is based on the energy consumption and data dissemination in WSNs. Direct diffusion [9] is a data centric routing approach for data aggregation in WSNs. Gradient broadcasting (GRAB) [10] is a general scheme for collecting information where target or the user collecting the information is fixed. GRAB aims at rich data delivery in large WSNs. Low-energy adaptive clustering hierarchy (LEACH) [11] is a clustering based protocol to collect data from WSN. LEACH is an energy conserving protocol based on the clustering for aggregating the data to the CHs. Geographical multicast routing protocol (GMR) [12] routes data through the shortest path, but the location update messages are individually forwarded by a mobile sink to sources whenever there is movement in each sink. Region based data dissemination scheme RBDD [13] performs local flooding within group region; based on the current location of the mobile sink. All the sinks receive data without any location updates. It supports mobility for sinks that move in or out of their region. RBDD offers improved data dissemination and is energy efficient scheme for WSNs. Trajectory and energy-based data dissemination protocol (TEDD) [14] combines the concepts of trajectory based forwarding with the power levels of SNs to calculate forwarding paths. When a SN receives a data packet, it decides (based on its location) whether the data packet should be forwarded. The data dissemination system adapts dynamically on the left over energy of the SNs. Due to the limited processing capacity and energy of SNs, TEDD is not quite easy to implement. Hierarchical cluster based data dissemination (HCDD) [15] scheme organizes SNs into a hierarchical structure, so that each SN has to locally exchange the information with its immediate neighboring nodes. HCDD builds single or multiple hierarchical structures to support multiple source nodes. HCDD avoids the increasing overhead of route discovery if the number of source nodes increase. HCDD is scalable for the large scale WSNs.

Virtual grid concept is simple to implement as compared to cluster based schemes for WSNs. It not only reduces storage requirements but also reduces the energy consumption of SNs. TTDD [2] is the scheme that works on the moving sink (the number of sinks may vary). This scheme utilizes the square virtual grid paths for data dissemination instead of using all the SNs of the whole sensor field. TTDD reduces energy consumption of the whole network. TTDD always maintains square virtual grid paths instead of using the shortest possible path, which is a diagonal path. Geographical and energy aware routing scheme (GEAR) [16] utilizes the geographical location information to route queries or data to any specific region in the wireless sensor field. If the locations of the sources are known, this scheme saves energy by limiting the flooding to that geographical region. Geographical adaptive fidelity scheme (GAF) [17] builds a geographical grid to turn off sensor nodes for minimizing the energy consumption. GAF grid is pre-determined and well synchronized in the complete wireless sensor filed, whose cell size is determined by the communication range of SNs. Distance vector multicast routing protocol (DVMRP) [18] supports data delivery from multiple sources to multiple destinations and faces similar challenges of TTDD. Energy Efficient Data Dissemination protocol (EEDD) [19], addresses the issues of target and inquirer mobility and energy conservation. EEDD improves the network lifetime by adopting a virtual-grid-based two-level architecture to schedule the activities of SNs. Data dissemination with ring based index [20] scheme collects, processes and stores sensed data at the nodes close to the detecting nodes. The location information of these storing nodes is pushed to some index nodes, which act as the rendezvous points for sinks and sources. A Diagonal-based TTDD (A-TTDD) [21] approach adopted the diagonal structures for data dissemination, so that energy consumption is reduced. Grid based data dissemination scheme (GBDD) [22], is a dual radio based grid construction scheme; which exploits dual radio mode of a sensor node to for data dissemination. Grid construction is initialized by the sink appearing in the sensor field when no valid grid is present. Any sink appearing during valid grid period shares existing grid and thus obviate the need to construct new grid. GBDD disseminates data diagonally across the grid using high power radio transmission mode in order to conserve energy.

3. Analytical Model

Assuming that N location aware SNs (coordinates known to each SN) are uniformly distributed over an area A (as shown in fig. 1). Sink & Source are mobile, whereas SNs are static. Sink and SNs have transmission range R. Sink sends a query message to all the SNs within its transmission radius (R) (as shown in fig. 4). There may be K_i moving sinks in the sensor field (where i=1, 2, 3, 4….).The sink moves with an average speed S. Each node transmits d data packets in time period T, of size $PackLEN$. If there are n SNs in a cell, then there will be $\sqrt{n}$ SNs on each side of the cell. Let the grid size be $\alpha \times \alpha = \alpha^2$, where α is the distance between the crossing points of the grid.

The sensor nodes can be arranged into a grid structure as shown in figure 1. Sink initiates the grid formation by broadcasting an election message and its coordinates. SNs in area πR^2around the sink will be receiving this message.

If only those sensor nodes which falls in radius greater than *R/2* from the sink respond to sinks message; then other SNs will be conserving their energy. The area ($A_{respond}$) in which SNs respond to sink's election message is given by Eq. (1) and is shown in figure 2. The area ($A_{respond}$) whose nodes will respond (as shown in fig. 2 is given by Eq. (1).

$$\begin{aligned} A_{respond} &= \pi(R)^2 - \pi(R/2)^2 \\ &= \pi\,[R^2 - R^2/4] \\ &= \pi\,[3R^2]/4 \\ &= \pi[\sqrt{3}/2\ R]^2 \end{aligned} \quad (1)$$

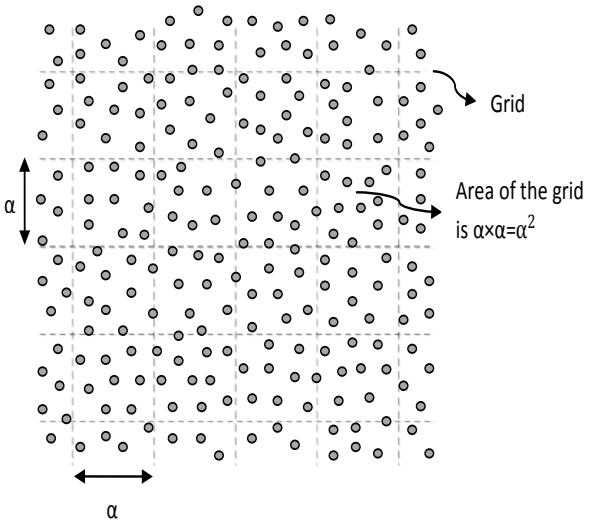


Fig. 1 Grid structure for area $\alpha\times\alpha=\alpha^2$.

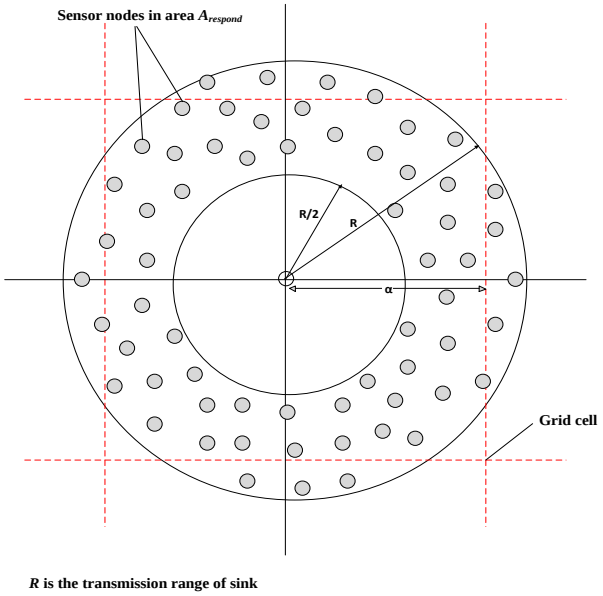


Fig. 2 Area ($A_{respond}$) in which SNs respond to sink's election message.

3.1 Energy Consumption

Assuming, initial energy of a SN is $E_{i,initial}$, $E_{i,sense}$ is per bit sensing energy consumed by a SN, $E_{i,Tx}$ is the transmission energy consumed per bit, $E_{i,Rx}$ is the receiving energy consumed per bit and $E_{i,process}$ is the processing energy consumed per bit by a SN, then the total energy consumed ($E_{i,Total}$) by a SN at particular instant of time is given by Eq. (2).

$$\begin{aligned} E_{i,Total} = {} & (b(E_{i,sense})) + ((p)(n)(E_{i,Rx})(PackLEN)) + \\ & ((m)(E_{i,Tx})(PackLEN)) + \\ & ((k)(E_{i,process})(b+((p)(n)(PackLEN))) \end{aligned} \quad (2)$$

SN which will act as dissemination nodes (DNs) will be responsible for transmitting the packets received from the non-dissemination nodes towards the sink. Assuming, $E_{i,DN}$ is the energy consumed by a DN at any particular time and $E_{i,LN}$ is the energy consumed by a non-dissemination node at any particular time. Energy consumption of $E_{i,DN}$ and $E_{i,LN}$ at particular instant can be derived from Eq.(3) and is given by Eq.(4) and Eq.(5). DN will be consuming energy in sensing, receiving from non-dissemination nodes, processing the received and sensed data and in transmission of packets towards the sink. Energy consumption of non-dissemination nodes is less as compared to DNs as they will only consume energy in sensing and transmitting packets to DN.

$$E_{i,DN} = E_{i,Total} \quad (3)$$

$$E_{i,LN} = (b(E_{i,sense})) + ((p)(E_{i,Tx})(PackLEN)) \quad (4)$$

Where, *b* is the number of bits sensed by the SN, *p* is the number of data packets sent by a SN to DN, *n* is the number of SNs from which DN is receiving packets, *m* is the number of data packets sent by the DN, *k* is the number of data packets processed by the DN and *PackLEN* is the data packet length.

3.2 Communication overhead

Assuming a rectangular sensor field of area *A* in which there are $n = \frac{N\alpha^2}{A}$, SNs in each cell and $\sqrt{n}$SNs on each side of cell. The data packet has a unit size and the messages have comparable size *L*. Assuming that there are *k* sinks in the sensor field moving with an average speed *S*. Sink receives *d* data packets from the source in the *T* time period. Further assuming that the sink traverses *m* cells, where the upper bound of *m* is $1+\frac{ST}{\alpha}$ and if *m*=1, then the sink is stationary. Sink updates its location *m* times as it traverses *m* cell and receives $\frac{d}{m}$ data packets between two successive locations.If sink updates its location by flooding a query locally to reach its nearby dissemination nodes only(as explained in section 4.4 for TBEE) then overhead for the query (without considering query aggregation) is *nL*, where *nL* is the local flooding

overhead. The overhead for *k* mobile sinks is then *km (nL).*

If WSN consists of *N* SNs, then for updating a mission additional overheads for TBEE (as explained in section 4) are *NL* and $\frac{4NL}{\sqrt{n}}$ (for grid formation). The total communication overheads for TBEE are given by Eq. (5)

$$CO_{TBEE} = NL + \frac{4NL}{\sqrt{n}} + kmnL \quad (5)$$

The total communication overheads of TTDD [2] is given by Eq. (6)

$$CO_{TTDD} = NL + \frac{4NL}{\sqrt{n}} + kmnL + kc(mL + d)\sqrt{2N} \quad (6)$$

The comparison of TBEE and TTDD in terms of communication overhead can be done considering a scenario where a sensor network consists of *N*=10,000 SNs, there are *n*=100 sensors in a TBEE grid cell. Suppose *c*=1 and *L*=1, to deliver *d*=100 data packets. For the stationery sinks, *m*=1 and suppose there are two sinks (k=2), then $\frac{CO_{TBEE}}{CO_{TTDD}}$ = 0.33. If sinks are mobile then ratio of communication overhead is, $\frac{CO_{TBEE}}{CO_{TTDD}} \rightarrow$ 0.4142, as m $\rightarrow \infty$. The above comparison shows that TBEE has less communication overheads as compared to TTDD, hence is more energy conserving.

4. TBEE: Tier Based Energy Efficient Protocol

TBEE has been designed for WSNs where the nodes know their respective coordinates. Sink initiates the process of grid formation by sending an election message. Along with election message sink sends the calculated coordinates of DPs, which are virtual cross-section points of the grid. Nodes on receiving election message calculate their respective distance from DP (to which they are closest) and respond to sink with their coordinates and their respective distance from the closest DP. Sink elects the nodes closet to DPs as dissemination nodes (DNs) by sending an appointment message. Later subsections and scenarios explain the working of TBEE.

4.1 Grid Construction

TBEE has been designed for WSNs where transmission range (*R*) of nodes and sink are same. All the nodes in the area πR^2 from a particular sink or other node will receive their transmission. Sink node calculates DP in all four direction at distance α ($\left(\frac{3R}{4}\right) < \alpha < R$). The location of DPs will be $(x \pm \alpha, y \pm \alpha)$.Grid formation is initialized by sink by broadcasting an election message along with its coordinates and mathematically calculated DPs (of all the four directions). SNs which are in the radius of *R*/8 of DP upon receiving this message; calculate their respective distance from DPs. SNs respond to sink by broadcasting their coordinates and the calculated distance from DPs. Sink elects dissemination nodes (in all four directions) which are closest to DPs by sending an appointment message along with coordinates of the nodes. If there are two or more SNs, who have same distance from the dissemination point then anyone of them can be elected as DN. Similarly, appointed DNs will further elect other DNs in entire sensor field. The appointment of new DN by another DN is shown in figure 3. SNs which fall in radius less than *R*/8 from DP upon receiving election message transmit their calculated distance from DP and their coordinates. The node (A) closest to DP is appointed as DN. A virtual grid is formed by appointment of DNs throughout the sensor field.

TBEE can construct grids of size α, which is much larger than the transmission radius *R*, by appointing intermediate DNs (IDN). The process is similar to appointing DNs by the sink. The Sink broadcasts an election message for the formation of IDNs. The SNs which fall in the radius greater than $\frac{3R}{4}$ and within the distance $\frac{R}{8}$ from the intermediate dissemination point (IDP) will respond to the message by sending their coordinates distance from the IDP (as shown in figure 4).

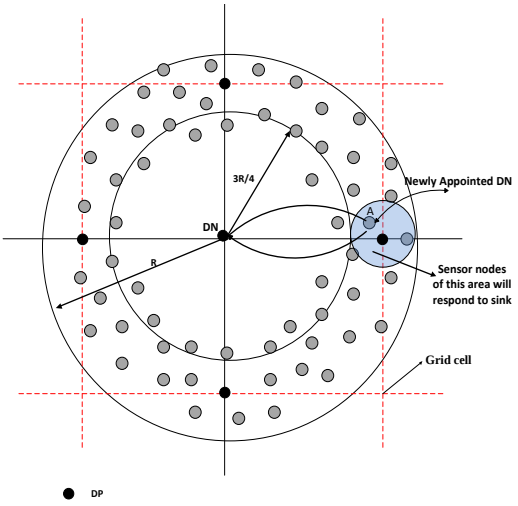


Fig. 3 Appointment of dissemination node during grid formation.

The Sink node will appoint an IDN within its transmission radius. Sink node sends the coordinates of DPs while appointing the IDNs. The SNs can communicate with sink either through DNs or IDNs. The IDNs calculate whether the dissemination point is within their transmission range or not. If DP is not in the transmission range of the IDN it

further appoints another IDNs till DN is not appointed. The Sink node will communicate to the DN through IDNs. While the grid is being constructed, the appointment of DNs in the direction where the DN has already been appointed is efficiently handled by TBEE. As shown in figure 5, A, B, C and D are the DNs formed by the DN, P. Now DN, B will broadcast the election message for further formation of new DNs. When this election message is received by the previously formed DN i.e. P, it sends a message to B notifying it it's coordinates (that it is the DN already appointed in that particular direction). SNs of remaining three directions will send their coordinates to B and B appoints the remaining DNs.

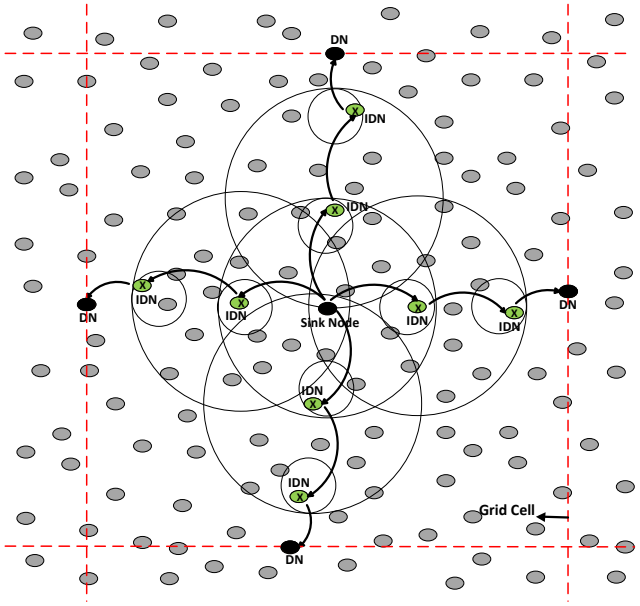


Fig.4 The formation of the grid when the value of α is much larger *R*.

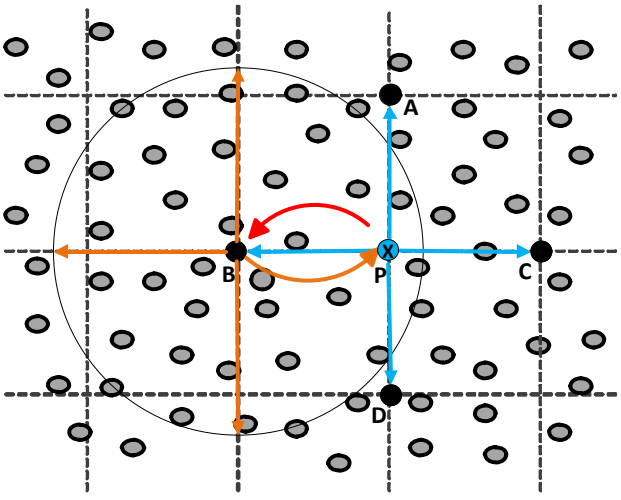


Fig. 5 Formation of the dissemination nodes.

4.2 Grid Termination

TBEE stops grid formation when the grid is formed for whole of the sensor field. As shown in figure 6, B is the DN and P is the DP. B broadcasts an election message for the further appointment of DNs. The nodes in the area $\frac{R}{8}$, from DP will respond to the election message by B. Since there are no SNs in that area of P (SNs which lie in the dark circled area have to respond to the B), B will not receive any message from that particular directions DP. The grid formation will be terminated in that particular direction by B but grid formation will continue in other directions. The SNs in that particular direction associate with B for sending their packets to sink.

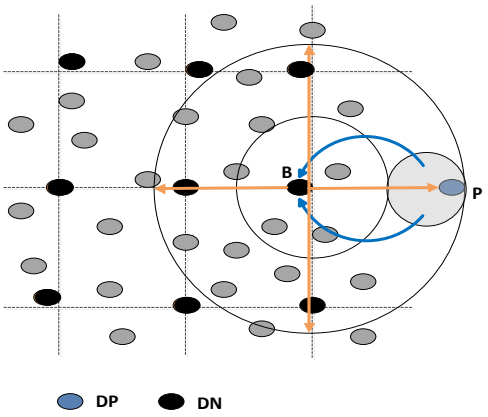


Fig. 6 Grid termination at the border.

4.3 Scenario 1: Grid maintenance with multiple sinks

The grid will be formed by TBEE as explained earlier but in case of a new sink that appears at any place in the network, the new sink will not construct its grid but will use the previous existing grid. The new sink will broadcast a message for the formation of its DNs. As shown in figure 7, DNs A, B, C and D formed by sink 1are in transmission range of sink 2, so upon receiving the message will respond to the election message of sink 2, intimating it about their status. Sink 2 will terminate the process of grid formation and will use grid formed by sink 1. Initial grid formed will be used by another new sinks for their data dissemination.

In case, when the DNs of the initial grid are not in transmission range of another sink then it will appoint its own DNs initially. New grid formation will continue in the direction where the DNs of previous grid are not in the transmission range of DNs of new grid. DNs of new grid will terminate the grid formation if it has at least two DNs of previous grid are in its transmission range. As shown in figure 8, E, F, G and H are the DNs appointed by sink 2.

When E broadcasts election message, A and B will respond to it by intimating their status. The grid formation is terminated by E. Similarly the grid formation will be terminated by other DNs formed by sink 2. Sink 2 will make use of grid formed by sink 1 through its DNs for data dissemination.

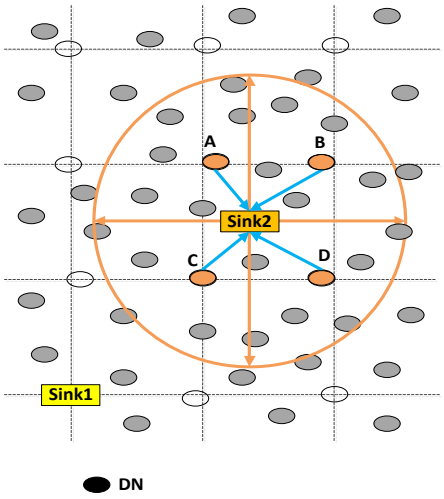


Fig. 7 Grid structure when multiple sinks exist.

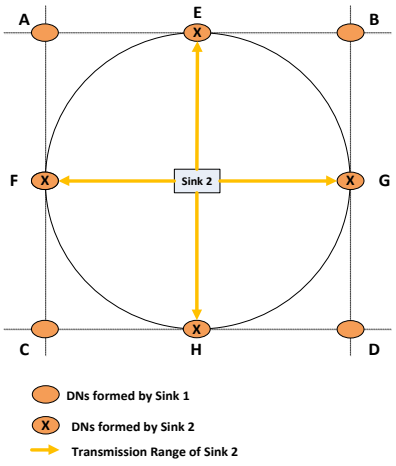


Fig. 8 Grid formation by sink 2.

4.4 Scenario 2: Grid maintenance when DNs reach threshold value of energy

The data dissemination will be through DNs; hence they will be consuming more energy as compared to other SNs (as given by Eq. (3) and Eq. (4)). In order to increase the life time of the network at DN upon reaching a threshold value of their energy (fifty percent of the initial energy), send an election message to SNs along with the coordinates of DP (since; it can be different than DN's coordinates) and coordinates of DNs which are in its transmission range. The nodes which are within the distance $\frac{R}{8}$ from the DP and at least two DNs are within their transmission range respond to this election message by sending their coordinates and the energy level to the electing DN. If the energy level of nodes is more than that of the electing DN, node with maximum energy level is appointed as DN. In case more than one node reports the same maximum energy level; appointing DN appoints any one of them as DN. The new DN will elect IDNs for the DNs which are not in transmission range of new DN. The appointment of new DN and IDNs are intimated to the sink by the new DN. If no SN responds to the election message of DN, then DN continues till it reaches another threshold level (seventy five percent of initial energy). DN if upon reaching this threshold level is unable to appoint new DN, it sends a re-election message to sink. A new grid formation is initiated by the sink.

4.5 Scenario 3: Movement of sinks

When sink is mobile it appoints SN nearest to it as DN. If the newly appointed DN is within transmission range of DNs(in all four directions) then it will not appoint any IDN otherwise it will appoint IDNs in the directions, whose DNs are not in its transmission range. As shown in figure 9, nearest node Sink 1is appointed as DN and is within the transmission range of its neighbouring DNs. Sink 1 moves from its initial location to new location; it comes under the transmission range of DN, A. Sink 1 will use A for data dissemination. If while moving sink is not in transmissions range of any DN it will appoint the IDNs in all four directions to communicate with the nearest DN. Similarly, Sink 2 appoints DN, C (as its nearest node). Now as it moves from its previous location to new location, it comes under the transmission range of DN, D.

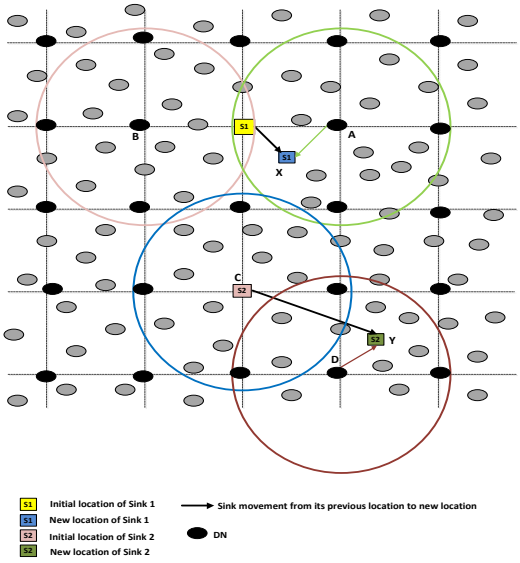


Fig. 9 Movement of sinks to new location.

4.6 Scenario 4: When source is mobile

TBEE can handle the mobility of source since every SN has a DN through which it transfers data diagonally towards the sink. As shown in figure 10 suppose the source is mobile the nodes sensing it will transmit the sensed data through their DNs towards the sink. In figure 10, when source moves to a new location all the data generated by the source will be transferred to sink through the DN, E of the grid.

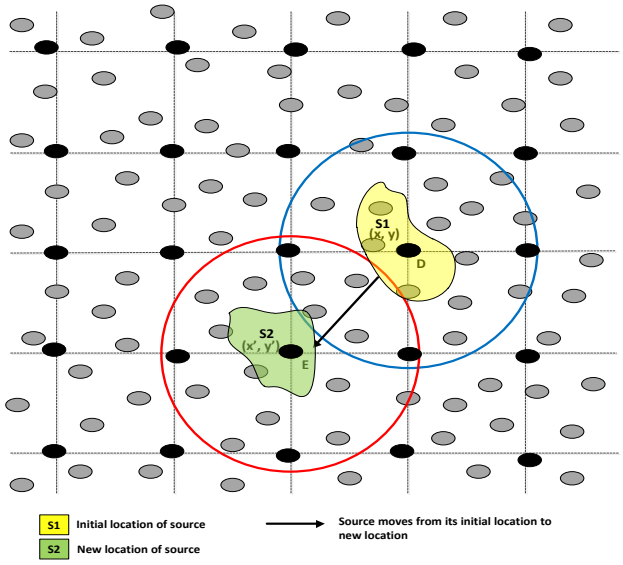


Fig. 10 Data dissemination by mobile source.

4.7 Scenario 5: Communication of SNs to the querying DN

TBEE is able to handle queries efficiently and avoids collisions so that retransmissions are avoided. Whenever sink wants information from some of the SNs the data is routed through a DN towards the sink. As shown in figure 11, SNs, S1 to S7 transmit their sensed data towards the sink through to DN, B.

SNs upon receiving query to send their sensed data wait for a guard time t_g before attempting to transmit anything. After the guard time expires or the channel is free, each SN will wait for a random listening time t_L before transmitting their data. The guard time t_g is to ensure that SNs reliably estimate the channel as either busy or idle. The additional random listening time t_L is to prevent nodes attempting to transmit their information at the same time. Suppose S1 sends RTS (Request to Send) message to B. The other SNs will be listening to the message will sense that the channel is busy. B will send CTS (Clear to Send) message to S1. Then, S1 will exchange the DATA & ACK (acknowledgement) packets with B; thereafter channel is free for further communication by other SNs. This above stated scheme of TBEE prevents collisions; hence energy consumption for retransmission is conserved.

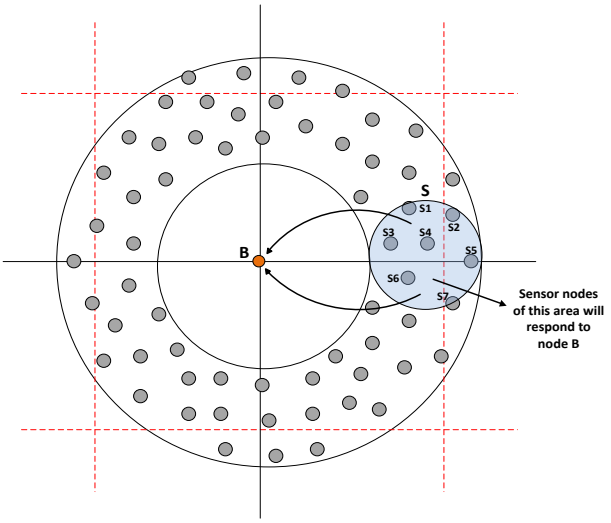


Fig. 11 Communication of SNs with node B.

5. Performance Evaluation & Simulation Results

Performance of TBEE was evaluated through simulations. Omnet++ an event based simulator was used for simulations. Simulation metric, parameters and evaluation methodology have been described in sub-section 5.1. The effect of various factors such as number sinks, varying number of sources and grid cell size on the performance of TBEE was evaluated. Performance of TBEE was compared with GBDD and TTDD.

5.1 Simulation metric, parameters and evaluation methodology

Same simulation parameters were taken for comparing TBEE with GBDD and TTDD. The SN's transmitting, receiving and idling consumption of power were taken as 0.66 W, 0.395 W and 0.035 W respectively. The simulations were performed with sensor field of 200 SNs, which are uniformly deployed in a 2000 X 2000 m^2 field. Packet length of query packet was considered as 36 bytes & each data packet was considered to be of 64 bytes. Various parameters used for simulation are given in table 1.

Two metrics were used to evaluate the performance of TBEE protocol. Our first evaluation metric is total energy consumption by SNs in transmitting and receiving queries and data. Energy consumption by nodes in idle state is not considered as it does not reflect energy consumption in

retrieval of data packets. Second, evaluation metric is the average delay which is defined as the average time taken by packets to reach sink from the source. It was averaged overall between source-sink pairs

Table 1. Simulation Parameters

Parameters	***Value***
Transmitting power of a SN	0.66 W
Receiving power of a SN	0.395 W
Idling consumption of a SN	0.035 W
Number of SNs	200
Area in which SNs are deployed	2000 X 2000 m^2
Query packet size	36 bytes
Data packet size	64 bytes
t_q	50 µs
t_L	100µs

5.2 Effect of number of sources and sinks on total energy consumption

The impact of variable number of sources and sink on total energy consumption has been evaluated for TBEE. There is varying number of sinks and sources to study and evaluate the effect on total energy consumption. Figure 12, shows the total energy consumption for the varying number of sinks with a single source. Results show that as the number of sinks increase the energy consumption also increases. Results of figure 13show the total energy consumed by the three protocols for the varying number of sinks with 8 sources. The total energy consumption increases as compared to the results of figure 12. The results of figure 12 and 13 show that the total energy consumption increases as the number of sources increase for all the three protocols but TBEE consumes less energy when compared with TTDD and GBDD.

5.3 Effect of number of sources and sinks on average delay

Figure 14 and figure 15 shows the average delay for varying number of sinks with single source and 8 sources respectively. It can be seen from the figures that average delay increases as the number of source increase. Average delay for GBDD is less as compared to TTDD and TBEE. This average delay is less for GBDD because SNs use high power radio for diagonal transmission of data towards the sink. The energy consumption of nodes is more if they use high powered radio for transmitting packets to a longer distance. There is an improvement by TBEE in terms of average delay when compared to TTDD.

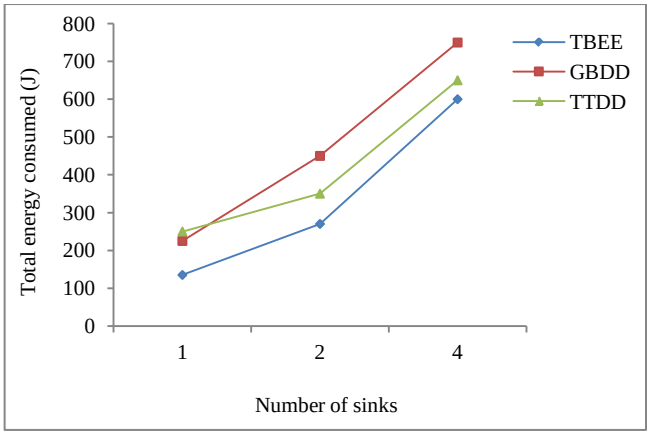


Fig.12 Total energy consumption for varying number of Sinks with single source.

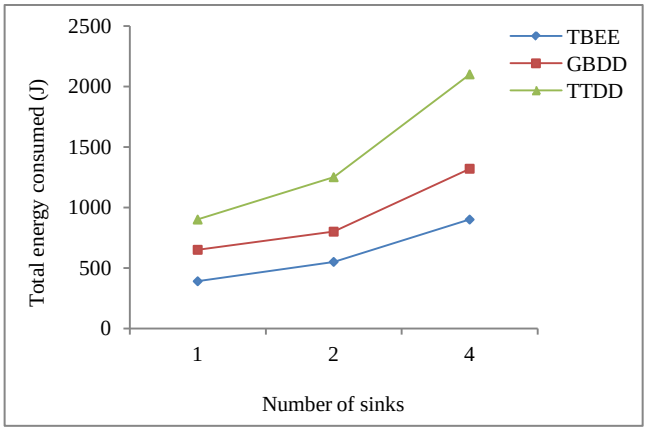


Fig.13 Total energy consumption for varying number of sinks with 8 sources.

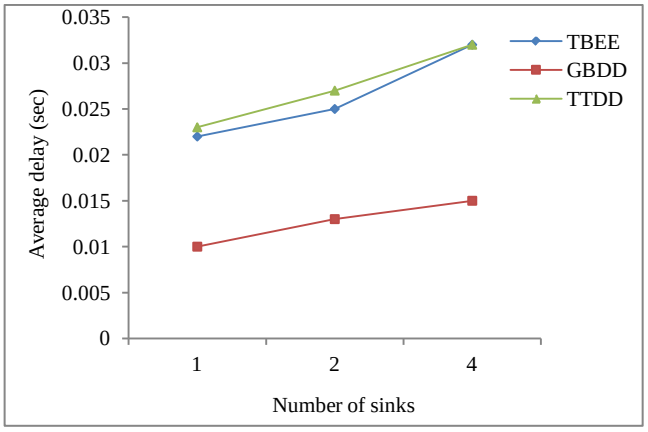


Fig.14 Average delay for varying number of sinks with single source

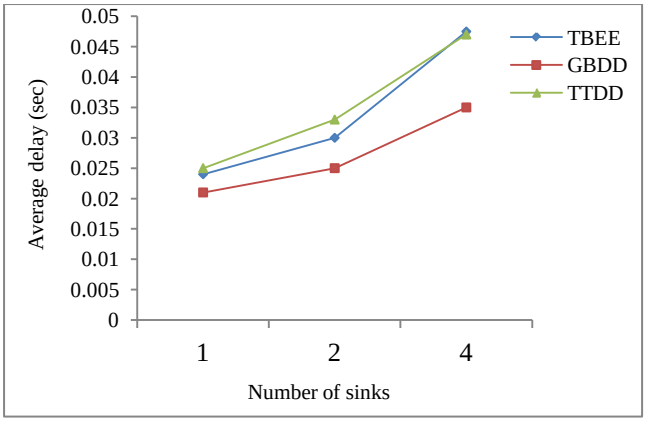


Fig.15 Average delay for varying number of sinks with 8 sources

5.4 Effect of Sink mobility on total energy consumption

Figure 16, shows the total energy consumption for mobile sink with varying speeds. If the speed of the sink is less the energy consumption is also less. As the speed increases the total energy consumption increases though energy consumed by TBEE is less as compared to TTDD and GBDD. This is because TBEE appoints less number of new DNs and makes use of the previous DNs.

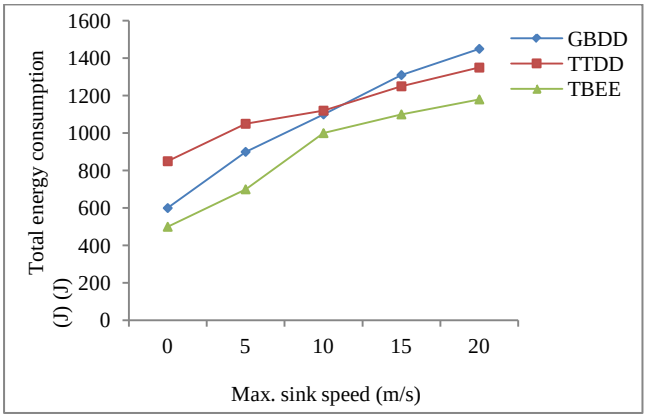


Fig.16 Total energy consumption for sink moving with varying speed

5.5 Effect of Cell Size α

The effect of cell size (α) on average energy consumption in the WSN for TBEE and TTDD is shown in figure 17. The results were obtained for1000 SNs deployed in the 6200 X 6200 m^2 sensor field. The nodes were placed evenly at 200 m distance with single source and sink. The cell size was varied from 400 m to 1800 m with an incremental step of 200m. Results show the average energy consumption of TBEE is less as compared to TTDD. The reduction in average energy consumption is more till the cell size is 1200 m, thereafter there is no significant reduction in the average energy consumption by TBEE. This results show that TBEE's energy consumption is effected by the grid cell size of the grid.

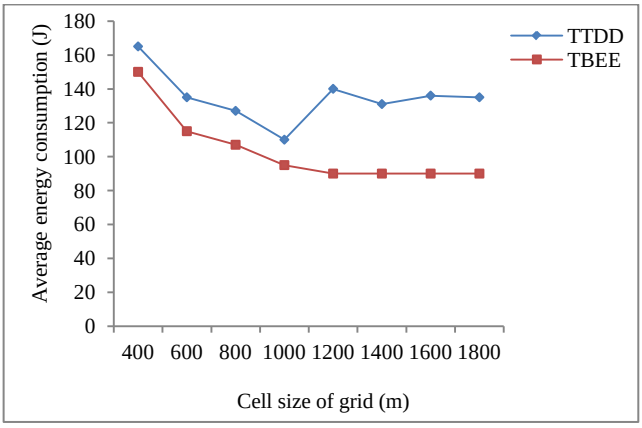


Fig. 17 Average energy consumption for varying cell size.

5.6 Effect of transmission radius for election message on TBEE.

The SNs which fall in a particular area from DP respond to election message. This reduces the overall energy consumption in a network. Results of figure 18 show average energy consumption of nodes for varying radius for communication in response to election message by sink. When the radius is $\frac{R}{2}$ from DP energy consumption is most but as the radius is reduced, the energy consumption also reduces because less nodes respond to the election message. Energy consumption is constant when the radius is further reduced from $\frac{R}{6}$, because the nodes are uniformly deployed and further reduction in radius does not reduce the number of nodes which respond to the election message.

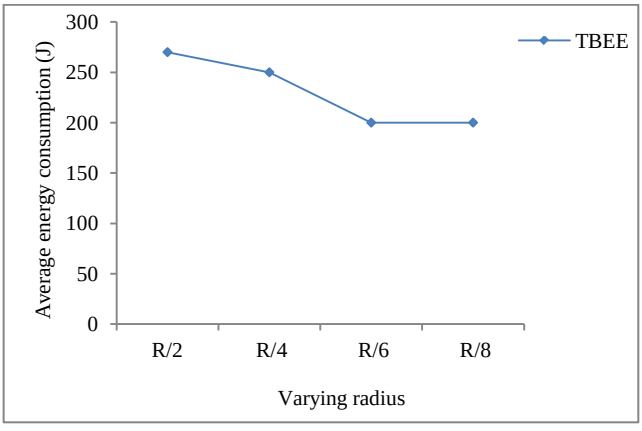


Fig.18 Average energy consumption of nodes for varying radius of communication to election message.

6. Conclusion

Virtual grid is quite useful and beneficial for the communication and data dissemination for the wireless sensor networks. Network lifetime can significantly be increased by reducing the transmission of packets. Our proposed protocol namely TBEE is energy efficient and is capable of handling sink and source mobility in wireless sensor networks. TBEE initially forms a virtual grid for the whole network. Grid formation is initiated by the sink and other sinks in the network use the previously constructed grid, which significantly reduces the energy consumption of the nodes. Mobility of the sinks and sources is efficiently managed by the message exchange and path discovery through the nearest dissemination nodes. Simulation results for TBEE show that when fewer nodes are located nearer to the dissemination points then the energy consumption is reduced. When the cell size of the grid is much larger as compared to transmission range of sensor nodes, the overall energy consumption in the network is reduced significantly by TBEE. Significant improvements in the simulated results are shown by TBEE when compared with TTDD and GBDD.

References

[1] F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “Wireless sensor networks: a survey”, Computer Networks, Vol. 8, No. 4, 2002, pp. 393–422.

[2] E. Ye, H. Luo, J. Cheng, S. Lu, and L. Zhang, “A two-tier data dissemination model for large scale wireless sensor networks”, in 8th ACM International Conference on Mobile Computing and Networking, 2002, No. 8, pp. 585–594.

[3] Zehua Zhou, Xiaojing Xang, Xin Wang and Jianping Pan, “An energy-efficient data-dissemination protocol in wireless sensor networks”, in International Symposium on on World of Wireless, Mobile and Multimedia Networks, 2006, pp.1–22.

[4] J. Cartigny, F. Ingelrest, D. Simplot-Ryl, and I. Stojmenovic, “Localized LMST and RNG based minimum energy broadcast protocols in ad hoc networks”, Ad Hoc Networks, Vol. 3, 2005 pp.1–16.

[5] Jia-Liang Lu, and F. Valois, “On the Data Dissemination in WSN”, in 3rd IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, 2007.

[6] V. Jolly and S. Latifi, "Comprehensive study of routing management in wireless sensor networks - part – 1”, in International Conference on Wireless Networks, 2006, pp. 37–44.

[7] S. Khan, Eui-Nam Huh, N. Imran, and Imran Rao, “A Membership Scheme for Gossip based Reliable Data Di ssemination in Ad-hoc WSNs”, in IEEE International Conference on Networking and Communications, 2008, pp. 107–111.

[8] W. Heinzelman, J. Kulik, and H. Balakrishnan, “Adaptive Protocols for Information Dissemination in Wireless Sensor Networks”, in 5th annual ACM/IEEE international conference on Mobile computing and networking, 1999, pp. 174–185.

[9] C. Intanagonwiwat, R. Govindan, and D. Estrin, “Directed Diffusion: A Scalable and Robust Communication Paradigm for Sensor Networks”, in ACM International Conference on Mobile Computing and Networking, 2000.

[10] F. Ye, S. Lu, and L. Zhang, “GRAdient Broadcast: A Robust, Long-lived Large Sensor Network”, http: //irl.cs.ucla.edu/papers/grab-tech-report.ps, 2001.

[11] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, “Energy efficient communication protocol for wireless micro sensor networks”, in 33rd Annual Hawaii International Conference on System Sciences, 2000, Vol. 2, pp. 1–10.

[12] J. Sanchez, P. Ruiz, J. Liu, and I. Stojmenovic, “Bandwidth-Efficient Geographic Multicast Routing for Wireless Sensor Networks”, IEEE Sensor Journal, Vol.7, No. 5, 2007, pp. 627–636.

[13] Hyungjoo Lee, Jeongcheol Lee, Sang-Ha Kim and Sungkee Noh, “Region Based Data Dissemination Scheme for Mobile Sink Groups in Wireless Sensor Networks”, in IEEE Global Telecommunications Conference, 2010, pp.1–5.

[14] M. Machado, O. Goussevskaia, R. Mini, C. Rezende, A. Loureiro, G. Mateus, and J. Nogueira, “Data dissemination in autonomic wireless sensor networks”, IEEE Journal on Selected Areas in Communications, Vol. 23, 2005, pp. 2305–2319.

[15] Ching-Ju Lin, Po-Lin Chou, and Cheng-Fu Chou, “HCDD: hierarchical cluster-based data dissemination in wireless sensor networks with mobile sink”, in ACM international conference on Wireless communications and mobile computing, 2006, pp. 1189–1194.

[16] Y. Yu, R. Govindan, and D. Estrin, “Geographical and Energy Aware Routing: A Recursive Data Dissemination Protocol for Wireless Sensor Networks”, Technical Report UCLA/CSD-TR-01-0023, UCLA Computer Science Dept., 2001.

[17] Y. Xu, J. Heidemann, and D. Estrin, “Geography Informed Energy Conservation for Ad Hoc Routing”, in 7th Annual ACM/IEEE International Conference on Mobile Computing and Networking, 2001.

[18] D. Waitzman, C. Partridge, and S. Deering, “Distance Vector Multicast Routing Protocol”, RFC 1075, 1988.

[19] Z. Zhou, X. Xang, X. Wang, and J. Pan, "An Energy-Efficient Data Dissemination Protocol in Wireless Sensor Networks", in International Symposium on a World of Wireless, Mobile and Multimedia Networks, 2006.

[20] W. Zhang, G.Cao, and L. Porta, "Data Dissemination with Ring-Based Index for Wireless Sensor Networks", IEEE Transactions on Mobile Computing, Vol. 6 No. 7, 2007, pp. 832–847.

[21] Bidi Ying, Huifang Chen, Wendao Zhao and Peiliang Qiu, “Intelligent Control and Automation”, in 6th World Congress on Intelligent Control and Automation, 2006, Vol. 1, pp. 257–260.

[22] Sharma, T.P., Joshi, R.C. and Misra, M., "GBDD: Grid Based Data Dissemination in Wireless Sensor Networks", in 16th International Conference on Advanced Computing and Communications,2008, pp. 234–240.

First Author Siddhartha Chauhan is Assistant Professor in Department of Computer Science and Engineering at National Institute of Technology, Hamirpur (H.P.) India. He did his masters in computer science and engineering from IIT, Roorkee. He has teaching and research experience of more than thirteen years. Presently he is pursuing his PhD. His research interests are routing, data dissemination and QoS in wireless sensor networks.

Second Author Lalit Kumar Awasthi is Professor in Department of Computer Science and Engineering at National Institute of Technology, Hamirpur (H.P.) India. He did his masters in computer science and engineering from IIT, Delhi and PhD. from IIT, Roorkee. He has more than twenty two years of experience in teaching and research. He has more than 120 research papers in journals and conferences. He is guiding many research scholars for their PhD. His research interests are check pointing in adhoc and mobile networks, wireless sensor networks and peer to peer networks.

Implementation of Variable Least Significant Bits Stegnography using DDDB Algorithm

Sahib Khan[1], Muhammad Haroon Yousaf[2] and Jamal Akram[3]

[1] Department of Telecommunication Engineering, University of Engineering and Technology, Taxila
Taxila, Punjab 47080, Pakistan

[2] Department of Computer Engineering, University of Engineering and Technology, Taxila
Taxila, Punjab 47080, Pakistan

[3] Department of Electrical Engineering, Federal Urdu University, Islamabad
Islamabad, 44000, Pakistan

Abstract

Nobody can deny the importance of secure communication. Different techniques are being utilized to achieve this task. Image Stegnography is one such method in which we hide data in an otherwise ordinary image. In this paper, a novel Stegnographic technique named as Variable Least Significant Bits Stegnography (VLSB) is proposed. To implement VLSB, we designed an algorithm named as Decreasing Distance Decreasing Bits Algorithm (DDDBA). In each test we performed, the data hiding capacity was always greater than 50 % (a barrier considered in image Stegnography), ranging up to 69 % with signal to noise ratio varying from 10 db to 5 db respectively. The DDDBA provides self-encryption mechanism in VLSB Stegnography, making the Steganalysis more difficult.

Keywords: *VLSB Stenography, DDDB Algorithm, Steganalysis, Key Size, Signal to Noise Ratio, Hiding Capacity.*

1. Introduction

The word 'Stegnography' literally means covered writing. It is a technique to camouflage the required information underneath an otherwise innocuous & routine data, in inconspicuous ways. Stegnography hides the covert information within the cover medium [1] making it difficult for anyone to detect even the presence of behind the scene secret message [2].

Stegnography is increasingly becoming popular especially in Defense Sector because of its distinctive features. In World War II, the first military use of Stegnography was seen and invisible inks were used for writing messages in between the lines of normal text message [3]. Germans in World War II used microdots. In this technology, the size of secret message containing photographs was reduced by a period. FBI director J. Edgar Hoover [4] called this technology "the enemy's master piece of espionage". With the development of digital images, new era of Stegnographic research started with multiple applications such as copyright protection, watermarking, fingerprinting, and Stenography [5, 7, 15 and 16]. Simmons' formulation of the Prisoners' Problem was itself an example of information hiding [8], [9].

Generally, information-hiding techniques are divided into two main categories: techniques in transform domain (e.g. Discrete Cosine Transform (DCT) [10] & Discrete Wavelet Domain [11] [12]), and techniques in time domain or spatial domain (e.g. LSB Stegnography, 4LSB Stegnography method [6]). 4LSB Stegnography has fixed data hiding capacity of 50% i.e. we need a cover file of almost double size as that of message file. To overcome this barrier, without compromising on security, a new technique called Variable Least Significant Bits (VLSB) Stegnography is devised. More details of data embedding and watermarking methods are available in [13]. Additional readings, software, and resources used in researching Stegnography and digital watermarking are available at [14].

2. VLSB Stegnography

Besides having a fixed limit of 50% data hiding capacity, 4LSB is relatively insecure as everyone can guess the position of actual data [17]. VLSB Stegnography, on the other hand has variable amount of data hidden in every individual pixel or group of pixels of the cover image. Cover image is divided in various groups of pixels, with each group being termed as a sector. The size of the sector is variable, ranging from the size of a complete cover file to that of a single pixel. Then, a specific number of bits "Bi", of each individual pixel of a sector, are used for data

hiding. The numbers of bits, used for substitution, varies from sector to sector according to a predefined algorithm.

The division of cover image into various sectors is the most crucial step for the implementation of VLSB Stegnography. The algorithm proposed should be capable of providing larger hiding capacity with least possible distortion. This will open a new research area for the researchers to play with VLSB Stegnography.

3. DDDB Algorithm

Decreasing Distance Decreasing Bits Algorithm (DDDBA) is a distance-based technique developed to implement VLSB Stegnography. First, the cover image is divided into various numbers of sectors on the basis of the distance of a pixel or group of pixels with respect to a specific reference point, usually the central pixel. The number of bits to be substituted in each pixel of a sector is decided on the basis of distance of that particular sector from the reference pixel. As the distance decreases, the number of bits to be embedded also decreases. That is why it is called Decreasing Distance Decreasing Bits Algorithm. The number of sectors "Ns" and the number of bits "Bi" to be substituted, play a vital role in determining hiding capacity, SNR/Distortion and key size. Large number of sectors results in small sector size, low distortion, large SNR and smaller hiding capacity and vice versa. By increasing the number of sectors to infinity, the sector size tends to zero and the whole cover file is treated as a single sector by the proposed algorithm; and for this particular case the VLSB Stegnography becomes equivalent to 4LSB Stegnography

2.1 Hiding Capacity of DDDB Algorithm

Decreasing According to DDDB Algorithm, the cover image is divided into "Ns" number of sectors, each of size "S_{zi}". Then "B_i" number of bits is hidden in each individual pixel of sector "Si". Therefore, the total number of bits "Di" hidden in sector "S_i" of size "S_{zi}" is given by

$$D_i = S_{zi} \times B_i \quad (1)$$

The total amount of data "D_{total}" hidden in the cover image can be calculated as:

$$D_{total} = \sum_{i=1}^{Ns} D_i \quad (2)$$

The hiding capacity "C" can be found by:

$$C = \frac{D_{total}}{B_{total}} \times 100 \quad (3)$$

$$C = \frac{\sum_{i=1}^{Ns} D_i}{N \times 8} \times 100 \quad (4)$$

$$C = \frac{\sum_{i=1}^{Ns} (S_{zi} \times B_i)}{N \times 8} \times 100 \quad (5)$$

Obviously, to get a data hiding capacity of more than 50% B_i should be greater than or equal to 4.

2.2 Key Size of DDDB Algorithm

As mentioned in the previous section, DDDB Algorithm divides the cover image into "Ns" number of sectors. Then each sector can be used to hide a number of bits "B_i" ranging from 0 to 8 ($0 \leq B_i \leq 8$) i.e. we have 9 possibilities for each sector. Therefore, the total possible ways (Key Size) to implement VLSB Stegnography using DDDB Algorithm is given by:

$$KeySize = N_s \times C_1^9 \quad \text{for } 0 \leq B_i \leq 8 \quad (6)$$

$$KeySize = N_s \times 9 \quad (7)$$

However, for $0 \leq B_i \leq 8$ the data hiding is smaller the 50%. To get a data hiding capacity which is greater than 50%, B_i should be greater than or equal to 4 i.e. $4 \leq B_i \leq 8$. For this range, we are having 5 different values of B_i so the Key Size for more than 50% data hiding capacity will be

$$KeySize = N_s \times C_1^5 \quad \text{for } 4 \leq B_i \leq 8 \quad (8)$$

$$KeySize = N_s \times 5 \quad (9)$$

Therefore, the capacity is increased at the cost of reduced Key Size.

2.3 SNR and PSNR of DDDB Algorithm

The quality of the stego-image is measured quantitatively by calculating signal to noise ratio (SNR) and peak signal to noise ratio (PSNR).

SNR and PSNR for a stego image are calculated in Decibels as:

$$SNR = -10Log\left[\frac{sum((Coverimage - Stegoimage)^2}{sum((Coverimage)^2)}\right]^{-1} \quad (10)$$

$$PSNR = -10Log\left[Mean((Coverimage - Stegoimage)^2)\right] \quad (11)$$

4. Implantation

In DDDB algorithm point, usually the centre pixel is selected as a reference. Then the maximum distance between the central pixel and border pixels is determined. The cover image is divided into a number of sector (Ns) each of Size (Sz). Each sector of cover image is assigned a specific number of bits "Bi" to be embedded in each pixel of that sector. The number of bits used for data hiding varies from sector to sector based on its distance from the central pixel. According to DDDB Algorithm, the number bits to be substituted decreases with decreasing distance of the pixel from the centre of the cover image. There are three types of distances; Euclidean, Chess Board and City Block. Each of the three can be used for implementing VLSB Stegnography using DDDB Algorithm. As shown in figure 1.

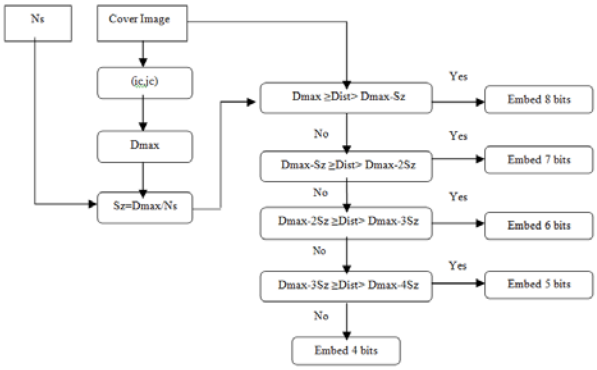


Fig. 1 Block Diagram of DDDB Algorithm

5. Experimental Results

Both the Qualitative, as well as the quantitative analysis of VLSB Stegnography using DDDBA was done, results were obtained & then analyzed. The experimental results using Euclidean, Chess Board and City Block distance are shown and compared in the following sections. However, the results obtained by using city block distance are not that much significant due to noticeable distortion and that too with less hiding capacity.

5.1 DDDB Algorithm with Euclidean Distance

Variable Least Significant Bits Stegnography implemented using Decreasing Distance Decreasing Bits Algorithm with Euclidean distance [18 and 19] and the resulted stego images for varying number sectors "Ns" and sector size "S_z" are obtained. The resulted stego images for Ns= 8, 16, 32, 64 and ∞ are shown here in figure 2 (a, b, c, d and e) respectively and the stego image obtained from 4SLB Stegnography is shown in figure 2(f).

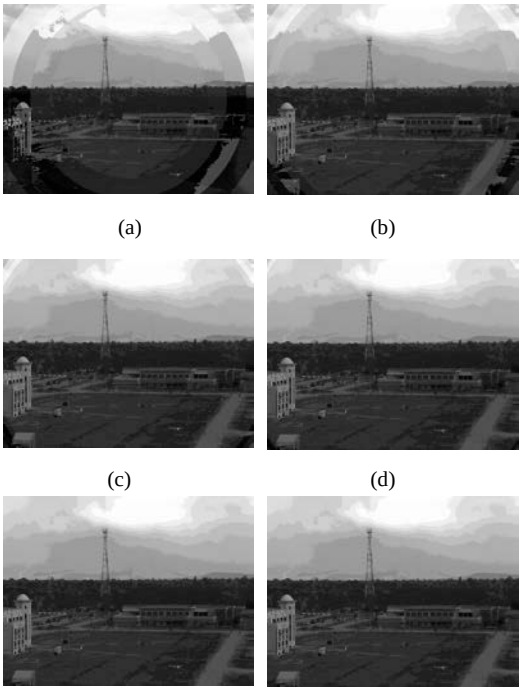


Fig. 2: (a)-(e) show five stego images obtained by implementing VLSB Stegnography using DDDB Algorithm for different number of sectors "Ns=8,16,32,64 and ∞"; (f) shows the stego image obtained by using 4LSB Stegnography

Quantitative data of hiding capacity, SNR and PSNR is shown in table 1. When we increase the number of sectors, both hiding capacity and distortion decreases. When the cover file is divided into infinite number of sectors, the VLSB Stegnography using devised Algorithm becomes equivalent to 4LSB Stegnography. The data hiding capacity and distortion created using both the techniques become equal. The result is shown in figure 2 (f, e).

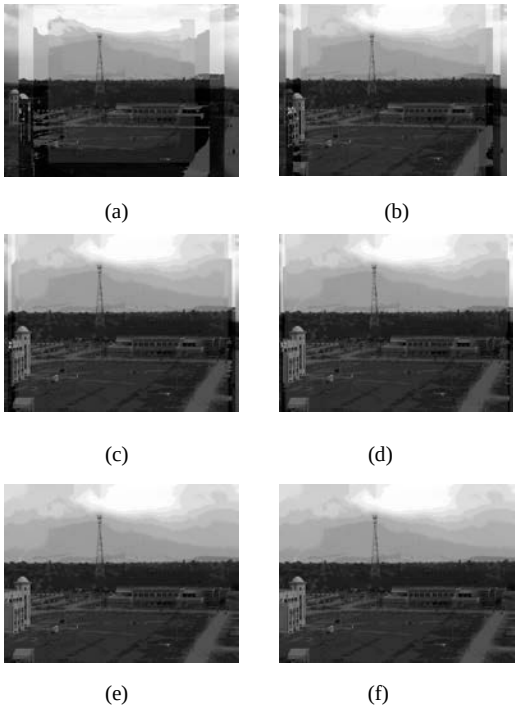
It is apparent from table 1 that hiding capacity of VLSB Stenography using DDDB algorithm is always higher than or equal to 4LSB Stegnography. Signal to noise ratio and peak signal to noise ratio are also affected by the number of sectors. SNR increases with increasing number of sector and vice versa.

Table 1: Hiding Capacity, SNR and PSNR of DDDB Algorithm with Euclidean Distance

Sr. No	Ns	Hiding Capacity	SNR	PSNR
1	8	64.6046	5.7902	-18.2743
2	16	53.8407	8.5163	-15.5482
3	32	50.8353	9.8855	-14.1790
4	64	50.1999	10.1694	-13.8951
5	Infinity	50.0000	10.2808	-13.7836
6	4LSB	50.0000	10.2808	-13.7836

5.2 DDDB Algorithm with Euclidean Distance

From Variable Least Significant Bits Stegnography is implemented with DDDB Algorithm using Chess Board distance [18] and results for varying number of sectors "Ns" and sector size "Sz" are obtained. The resulted stego images for Ns= 8, 16, 32, 64 and ∞ are shown in figure 3 (a, b, c, d and e) respectively. The stego image developed with 4LSB Stegnography is shown in figure 3(f). Quantitative data of hiding capacity, SNR and PSNR is shown in table 2.



(a) (b) (c) (d) (e) (f)

Fig. 3: (a)-(e) show five stego images obtained by implementing VLSB Stegnography using DDDB Algorithm with Chess Board Distance for Ns=8,16,32,64 and ∞ respectively; (f) shows the stego image obtained by using 4LSB Stegnography

Table 2: Capacity, NSR and PSNR of DDDB Algorithm with Chess Board Distance

Sr. No	Ns	Capacity	SNR	PSNR
1	8	69.0926	5.1915	-18.8729
2	16	57.8509	7.8138	-16.2507
3	32	53.8966	8.9079	-15.1565
4	64	51.9387	9.4960	-14.5685
5	Infinity	50.0000	10.2808	-13.7836
6	4LSB	50.0000	10.2808	-13.7836

While increasing the number of sectors, both hiding capacity and distortion decreases. When the cover file is divided into infinite number of sectors, the VLSB Stegnography using devised Algorithm becomes equivalent to 4LSB Stegnography. The data hiding capacity and distortion created using both techniques become equal. The result is shown in figure 3 (a-f).

6. Conclusions

In this paper, VLSB Stegnography using DDDB Algorithm is being proposed to achieve image Stegnography with desired results. Due to variable bits substitution, variable amount of data is hidden in different sectors of cover image depending upon the distance from the reference point. More data is hidden in border pixels, creating more distortion at the boundary of stego image. Due to eyesight limitation and false perception, the distortion at the border creates no significant effect if the number of sectors is large. When the number of sectors is made equal to infinity, VLSB Stegnography using DDDB Algorithm and 4LSB Stegnography techniques became equivalent as shown in figure 2(e and f), figure 3(e and f), table 1 and table 2. Moreover, for the same number of sectors, greater hiding capacity can be achieved using chessboard distance as compared to the Euclidean distance in which lesser hiding capacity is available, though with minimum distortion. However, both can effectively be used for hiding capacity of 50% and more.

References

[1] S. Dumitrescu, W.X.Wu and N. Memon (2002), "On steganalysis of random LSB embedding in continuous-tone images", Proc. International Conference on Image Processing, Rochester, NY, pp. 641-644.

[2] S.K. Moon and R.S. Kawitkar (2007)," Data Security using Data Hiding", International Conference on Computational Intelligence and Multimedia Applications, pp.247-251.

[3] D. Kahn and Macmillan (1967),"the Codebreakers", New York.

[4] Beenish Mehboob and Rashid Aziz Faruqui (2008)," A Stegnography Implementation", IEEE.

[5] Kafa. Rabah (2004), "Steganography - The Art of Hiding Data", Information technology Journal 3 (3).

[6] J. Fridrich, M. Goljan, and R.Du (2001),"Detecting LSB Stegnography in Color and Gray –Scale Images", Magazine of IEEE Multimedia, Special Issue on Security, October-November issue, pp.22-28.

[7] T. Cedric, R. Adi and I. Mcloughlin (2000)," Data concealment in audio using a nonlinear frequency distribution of PRBS coded data and frequency-domain LSB insertion", Proc. IEEE International Conference on Electrical and Electronic Technology, Kuala Lumpur, Malaysia, pp. 275-278.

[8] Gustavus J. Simmons (1998), "How to insure that data acquired to verify treaty compliance are trustworthy," *Proc. IEEE*, vol. 76, p. 5.

[9] Gustavus J. Simmons (May,1998) , "The history of subliminal channels," IEEE Journal on selected areas in communication, vol. 16, no. 4, pp. 452–462.
[10]Chiou-Ting Hsu, Ja-Ling Wu (January, 1999)," Hidden Digital Watermarks In Images", IEEE Transaction on Image Processing, 8(1): 56-58.
[11]Nedeljko Cvejic, Tapio Seppanen (Oct 2002)," A Wavelet Domain LSB Insertion Algorithm For High Capacity Audio Steganography", Digital Signal Processing, workshop 2002 and the 2nd signal processing education workshop. Proceedings of 2002 IEEE, 10th, 13-16, Pages: 53-55.
[12]NI Ronggong, RUAN Qiuqi (Oct 2002),"Embedding Information into Color Images Using Wavelet, TENCON'02", Proceedings, 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, Volume:1, 28-31, pages: 589-601..
[13]M.D. Swanson, M. Kobayashi, A.H. Tewfii (June, 1998), "Multimedia Data-Embedding and Watermarking Technologies", *Proc. of the IEEE*, vol. 86, no. 6, pp. 1064-1087.
[14]M.D. Swanson, B. Zhy and A.H. Tewfik (1996), "Transparent robust image watermarking", *Proc.* IEEE *International Conference on Image Processing* (ICIP96), Piscataway, NJ. IEEE Press, vol. 3
[15]S. K. Moon, V. N. Vasnik, "Application of steganography on image file", National conference on Recent trends in Electronics, pp. 179-185.
[16]T. Morkel, J. H. P. Eloff, M. S. Olivier,"An Overview of Image Steganography", Information and Computer Security Architecture (ICSA) Research Group, Department of Computer Science, University of Pretoria, SA.
[17]Sahib Khan and M. Haroon Yousaf, " Variable Least Significant Bits Stegnography", Accepted in IJCSI, 2011 (To be published)
[18]R.C. Gonzalez, R.E. Woods. *Digital Image Processing.* 2nd Ed, Prentice-Hall, Upper Saddle River, New Jersey, 2002.
[19] xlinux.nist.gov/dads/HTML/euclidndstnc.html

Engr. Sahib Khan is pursuing M.Sc in Telecommunication Engineering at department of Telecommunication Engineering, Faculty of Telecommunication and Information Engineering, University of Engineering and Technology Taxila. He has B.Sc Telecommunication Engineering from N-W.F.P University of Engineering and Technology Peshawar, Pakistan. He is serving as a Lecturer and Course Coordinator at department of Electrical and Computer Engineering, Kohat University of Science and Technology Kohat, Pakistan. His areas of interest are Digital Image Processing and Data Authentication and Information Security.

Engr. Muhammad Haroon Yousaf is pursuing Ph. D. in Computer Engineering at University of Engineering & Technology Taxila, Pakistan. He got M.Sc. and B.Sc. in Computer Engineering in the years 2007 and 2005 respectively. He currently holds the position of Assist Prof in the Faculty of Telecommunication & Information Engineering. His research interests include gesture and activity recognition, human computer interaction, computer vision and machine learning. He, along with his students, is investigating face recognition based solution for automated attendance management system.

Engr. Jamal Akram has done BSc Electrical Engg with Honours from University of Engg & Technology, Peshawar in 2000. He did his Masters in Computer Engg from University of Engg & Technology, Taxila in the year 2008. He has worked for more than 10 years in National Fertilizers Corporation & Pakistan Telecommunications Company Ltd. His areas of interest include Image Processing, Information Security, computer-vision and antenna design. Currently he is faculty member of Federal Urdu University Islamabad. He is also a member of IEEE.

Voice Recognition using HMM with MFCC for Secure ATM

Shumaila Iqbal[1], Tahira Mahboob[2] and Malik Sikandar Hayat Khiyal[3]

[1] *Software Engineering, Fatima Jinnah Women University, Rawalpindi, Pakistan*

[2] *Software Engineering, Fatima Jinnah Women University, Rawalpindi, Pakistan*

[3] *Software Engineering, Fatima Jinnah Women University, Rawalpindi, Pakistan*

Abstract

Security is an essential part of human life. In this era security is a huge issue that is reliable and efficient if it is unique by any mean. Voice recognition is one of the security measures that are used to provide protection to human's computerized and electronic belongings by his voice. In this paper voice sample is observed with MFCC for extracting acoustic features and then used to trained HMM parameters through forward backward algorithm which lies under HMM and finally the computed log likelihood from training is stored to database. It will recognize the speaker by comparing the log value from the database against the PIN code. It is implemented in Matlab 7.0 environment and showing 86.67% results as correct acceptance and correct rejections with the error rate of 13.33%.

Keywords: *Voice recognition, Mel Frequency Cepstral coefficients, Hidden Markov Modeling, Forward Backward algorithm, Fast Fourier Transform, Discrete Cosine Transform, K-mean algorithm.*

1. Introduction

A computer system that automatically identifies and verifies the person by capturing the voice from a source like microphone is known as voice recognition. Voice recognition is one of the terms of biometric technology. It uses to provide any authentication to any system on the basis of acoustic features of voice instead of images. The behavioral aspect of human voice is used for identification by converting a spoken phrase from analog to digital format, and extracting unique vocal characteristics, such as pitch, frequency, tone and cadence to establish a speaker model or voice sample. In voice recognition, enrollment and verification processes are involved. Enrollment process describes the registration of speaker by training his voice features [1] [2]. And verification contains to verify the speaker by comparing his current voice features to pre stored features of voice. In real time, the verification process splits into two mechanisms. It first compares the unknown speaker to the pre stored database of known speakers on the basis of 1:N. and then it make decision of speaker to the exact match of 1:1. Where the one voice sample finally matched to only 1 template stored in the database [3]. Voice recognition has two categories text dependent and text independent. Text dependent voice recognition identifies the speaker against the phrase that was given to him at the time of enrollment. Text independent voice recognition identifies the speaker irrespective of what he is saying. This method is very often use in voice recognition as it require very little computations but need more cooperation of speakers. In this case the text in verification phase is different than in training or enrolment phase [2] [4].

In Early research Shi-Huang Chen and Yu-Ren Luo presents in [5] the MFCC as to extract features and trained and recognized using SVM. They defined the MFCC as the unique and reliable feature extraction technique. That was used to find the most usable features in detailed form. In recognition phase SVM (super Vector Machine) technique based on two class classifiers by defining the decision in binary form was introduced. It discriminate claimed speaker and imposter by +1 and -1 by maximizing the margins or minimizing the structural risks. It shows results averaged to 95.1% with ERR of 0.0%. That was considered as the best results under 22nd order of MFCC. In another research [6], Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi also extracted the voice features using MFCC that was trained and recognized using DTW. DTW (dynamic time Wrapping) a non linear sequence alignment is another technique that is used for recognition process. They find it best for time sequence between two speeches. Here the optimal wrapping path is achieved by wrapping

the time distance between two signals. One other research [7] has done by Ibrahim Patel and Dr. Y. Srinivas Rao. They represent the voice recognition with improvement of MFCC with frequency decomposition technique. They introduces sub band coding in their research. The integration of MFCC with sub band coding increases its efficincy and accurate classification as compared to MFCC separately. These two features of MFCC and integrated sub band decomposition with MFCC are used in HMM to train and recognize the speaker.

This paper is attentive of providing a security by developing a voice recognition system to secure the ATM (automatic transaction machine) using HMM with MFCC. The usage of MFCC for extracting voice features and HMM for recognition provides a 2D security to the ATM in real time scenario. MFCC is used to describe the acoustic features of speaker's voice. HMM forward backward estimation technique is used to train these features into the HMM parameters and used to find the log likelihood of entire voice. In recognition HMM is used to compare log likelihood to the pre-stored value and intended to recognize the speaker. If the log likelihood is matched then it is granted otherwise failed to use ATM system. The rest of the paper is alienated as. Section 2 demonstrates the features to be extracted through MFCC. Section 3 describes the use of HMM. Section 4 demonstrates the methodology of being using these techniques. Section 5 consists of the experimental results. Finally section 6 contains the principle conclusion.

2. Mel Frequency Cepstral Coefficients

MFCC is used to extract the unique features of human voice. It represents the short term power spectrum of human voice. It is used to calculate the coefficients that represent the frequency Cepstral these coefficients are based on the linear cosine transform of the log power spectrum on the nonlinear Mel scale of frequency. In Mel scale the frequency bands are equally spaced that approximates the human voice more accurate. Equation (1) is used to convert the normal frequency to the Mel scale the formula is used as

$$m=2595 \log_{10} (1+f/ 700) \quad (1)$$

Mel scale and normal frequency scale is referenced by defining the pitch of 1000 Mel to a 1000 Hz tones, 40 db above the listener's threshold. Mel frequency are equally spaced on the Mel scale and are applied to linear space filters below 1000 Hz to linearized the Mel scale values and logarithmically spaced filter above 1000 Hz to find the log power of Mel scaled signal [8] [9]. Mel frequency wrapping is the better representation of voice. Voice features are represented in MFCC by dividing the voice signal into frames and windowing them then taking the Fourier transform of a windowing signal. Mel scale frequencies are obtained by applying the Mel filter or triangular band pass filter to the transformed signal. Finally transformation to the discrete form by applying DCT presents the Mel Cepstral Coefficients as acoustic features of human voice.

3. Hidden Markov Modeling

HMM is defined as a finite state machine with fix number of states. It is statistical processes to characterize the spectral properties of voice signal. It has two types of probabilities. There should be a set of observation or states and there should be a certain state transitions, which will define that model at the given state in a certain time
In hidden markov model the states are not visible directly they are hidden but the output is visible which is dependent on the states. Output is generated by probability distribution over the states. It gives the information about the sequence of states but the parameters of states are still hidden. HMM can be characterized by following when its observations are discrete:

- N is number of states in given model, these states are hidden in model.
- M is the number of distinct observation symbols correspond to the physical output of the certain model.
- A is a state transition probability distribution defined by NxN matrix as shown in equation (2).

$$A= \{a_{ij}\}$$
$$a_{ij} = p\{q_{t+1} = j | q_t = i\}, \quad 1 \le i, j \le N,$$
$$\sum_{j=1}^{N} a_{ij} = 1, \quad 1 \le i \le N \quad (2)$$

  Where q_t occupies the current state. Transition probabilities should meet the stochastic limitations

- B is observational symbol probability distribution matrix (3) defined by NxM matrix equation comprises

$$b_j(k)= p\{o_t=v_k | q_t=j\}, \quad 1<=j<=N , 1<=k<=M$$
$$\sum_{k=1}^{M} b_j(k) = 1, \quad 1 \le j \le N \quad (3)$$

  Where V_k represents the K^{th} observation symbol in the alphabet, and O_t the current parameter vector. It must follow the stochastic limitations
- π is a initial state distribution matrix (4) defined by Nx1.

$$\pi= \{\pi_i\}$$
$$\pi_i = p\{q_1 = i\}, \quad 1 \le i \le N \quad (4)$$

By defining the N, M, A, B, and π, HMM can give the observation sequence for entire model as λ= (A, B, π) which specify the complete parameter set of model [10].
HMM define forward backward estimation algorithm to train its parameters to find log likelihood of voice sample. Segmental k mean algorithm is used to generate the code book of entire features of voice sample.
Forward backward algorithm is used to estimate the unidentified parameters of HMM. It is used to compute the maximum likelihoods and posterior mode estimate for the parameters for HMM in training process. It is also known as Baum Welch algorithm. It computes the $P(X_k \mid o_{1:t})$Posterior marginal or distribution. For all hidden state variables $X_k \in \{X_1, \ldots, X_t\}$. By given a set of observations as $o_{1:t} := o_1, \ldots, o_t$.
This inference task is commonly known as smoothing [10] [11]. This algorithm uses the concept of dynamic programming to compute the required values for the posterior margins efficiently in two processes first doing the forward estimations and then backward estimation.
Segmental K-mean algorithm is used to clustering the observations into the k partitions. It is the variation of EM (expectation-maximization) algorithm. That is used to determine the k-means of data distributed by Gaussian distribution. Its objective is to minimize (5) the intra-cluster variance or squared error function.

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} |x_j - \mu_i|^2 \qquad (5)$$

Here, k is the no of clusters and μ define the centroids of mean point of all points of the input vector. K-mean algorithm is used to first partition the input vector into k initial sets by random selection or by using heuristic data. It defines two steps to precede k-mean algorithm. Each observation is assigned to the cluster with the closest mean. And then calculate the new means to be centriod of observation in each cluster by associating each observation with the closest centroids it construct the new partition , the centroids are recalculated for new cluster until it convergence or observations are no longer remains to clustering .
It converges extremely fast in practice and return the best clustering found by executing several iterations. Its final solution depends on the initial set of clusters [12] [13]. For this, the number of clusters k must be defined to find otherwise it gives eccentric results.

4. Proposed Methodology

The methodology proposed in this research paper consists of two techniques MFCC and HMM. MFCC is used to extract the voice features from the voice sample. And HMM is used to recognize the speaker on the basis of extracted features. For this it first train the extracted features in the format of HMM parameters and to find the log value of the entire voice for recognition. Forward backward estimation technique is used to train the extracted features and find its parameters. This section explains the methodology step by step by explaining (Figure 1) the two techniques MFCC and HMM.

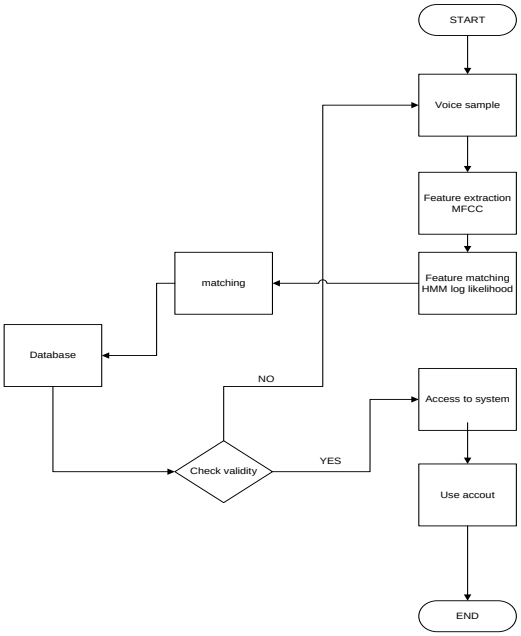


Fig .1 Proposed System

4.1 Voice Processing

The initial step to develop the proposed research is obtaining the voice sample.Voice sample is taken from the microphone by the speaker. It is digitalized by 8 KHz sampling rate for 2 seconds. Pre-emphasizing the signal make it to normalize. The pre-emphasizing (6) is done to balance the high frequency part of human voice that was covered up when he produce sound. It is also used to increase the high frequency formants in the speech.

$$X2(n) = X(n) - a*X(n-1) \qquad (6)$$

Where the value of a is between 0.9 and 1. Z transform (7) of the filter.

$$H(z) = 1 - a*z^{1} \qquad (7)$$

This pre-emphasized voice sample (Figure 2) is then stored to a wav file that is used for leading process.

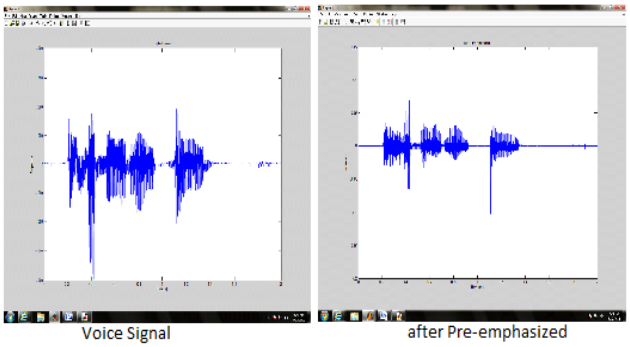


Fig. 2 voice processing

4.2 Features Extraction

In second step the obtained voice sample is used to frame in MFCC implementation. The pre emphasized voice signal is framed in order to get the stationary part of speech. The speech signal is divided into frames of 30~20 ms with optional overlap of 1/3~1/2 of frame size. With reference to sampling point the frame size usually maintained as the power of two to exploit the FFT. But if it is not then zero padding is done to the nearest length of power of two. Zero padding is used to extend the signal by adding zeros and by increasing its length N to M where M>N. The main purpose of this is to make the signal to the nearest length of power of two and make it feasible for FFT. In this proposed methodology, 256 sample points per frame (Figure 3) and 156 overlap frames are defined.

Framed speech signal is then multiplied with the hamming window in order to remove the discontinuities in the signal. Hamming window returns (8) the symmetric points of integral values framed signal into the column vector w.

$$w(n, \alpha) = (1 - \alpha) - \alpha \cos(2\alpha n/(N-1)) 0 \leqq n \leqq N-1 \quad (8)$$

Where α shows different curves of hamming window. Its value usually as 0~ 0.5 and window length is L=N. it is obtained (Figure 3) by multiplying each frame to the hamming window.

After windowing FFT is applied to convert (9) the signal into frequency domain from time domain and also used to obtain (Figure 3) the magnitude frequency response of each frame. In doing so it is assumed that the signal in frames in periodic and continuous when wrapping around. In the opposite case of this, there are some discontinuities at the frames start and end points that causes detrimental effects in frequency response. This can be overcome by multiplying each frame by the hamming widow that will help to remove discontinuities at the start and end points of frames.

$$y = \text{fft}(b) \quad (9)$$

Where b is the windowed form of signal.

In Mel filter or triangular band pass filter the magnitude of frequency response is multiplied with the 40 number of triangular band pass filters in order to obtain (Figure 3) the log energy of triangular band pass filter on Mel scale. These filters are equally spaced on the Mel scale (10) and use to calculate the linear frequency.

$$m = 2595 \log_{10}(1 + f/700) \quad (10)$$

The frequency response on Mel scale is reflecting the similar effect of human subjective auditory perception. Triangular band pass filter is used to flatten the magnitude spectrum and to reduce the size of the features occupied. Frequency wrapping (Figure 3) is applied here to keep the useful informational part of the Mel.

At the end by applying DCT cepstral features of voice signal are obtained (11). It is used to convert the log Mel scale cepstrum into time domain from frequency domain (Figure 3).

$$y(k) = w(k) \sum_{n=1}^{N} x(n) \cos\left(\frac{\pi(2n-1)(k-1)}{2N}\right) \quad k = 1, 2, \ldots N \quad (11)$$

Where N is the length of the computed Mel frequencies. The series starts from n and k=1, because MATLAB vectors starts from 1 instead of 0. The result is known as MFCC. These are the 40 acoustic features of human voice that are used to recognize the person depending upon the filter to be applied.

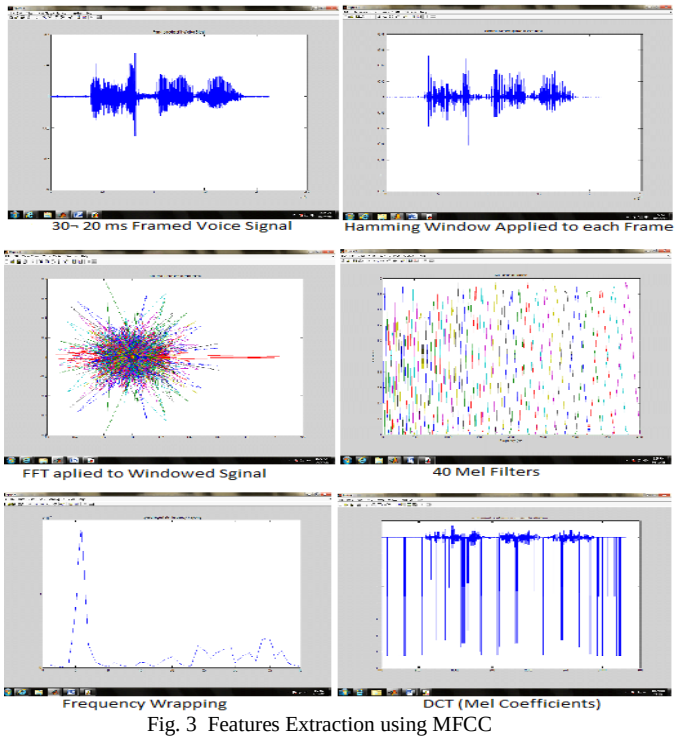


Fig. 3 Features Extraction using MFCC

Log energy is also an important factor of human voice to be recognized. It is computed by obtaining the frame energy of voice signal after framing. It is used to obtain defined number of coefficients of MFCC. Here it calculated the 15 number of coefficients of human voice.

Delta is the first order derivative of original cepstrum. It helps to make speech signal dynamic. There is also 15 coefficients are calculated as delta coefficients. It is the first order derivative of MFCC coefficients.

4.3 HMM Training

For HMM recognition the extracted feature vectors of MFCC are trained into HMM. The training is done in two steps as

- HMM code book
- HMM training by forward backward re-estimation algorithm

First the Code book contains the cluster number specifies to each observation vector, which is obtained by applying the K-mean algorithm. It is used to set the centroids of the observation vector. The observation vectors are represented in the form of matrix Y, and K is the desired number of clusters which are defines as 45; it is used to cluster the featured data Y by random selection. By clustering the model, it returns the centroids, one for each of the cluster k and refers to the cluster number or centroid index of centroid closest to it. K-mean algorithm tries to minimize the distortion that is defined as the sum of squared distances between each observation vector and its dominating centroids. Squared Euclidean distance is ordinary distance between two points which one can measure from ruler. It can be proved by repeated application of Pythagorean formula. In this research study, Euclidean distance (12) is used to find the distance between observation vector and its cluster centroids

$$\|Y-Yc\| = \text{sqrt}(\| Y \|^2 + \| Yc \|^2 - 2* Y. Yc) \quad (12)$$

After clustering the training to HMM parameters begins by applying Forward-backward algorithm. It uses the principle of Maximum likelihood estimation. It returns the state transition matrix A, observation probability matrix B, and the initial state probability vector π on the basis of defined states as 10 and codebook vectors. In this phase the observation vectors being trained in the form of HMM parameters and resulted as the log likelihood of entire voice. This log likelihood is used to store in speaker's database for recognition in real time.

4.4 HMM Recognition

HMM recognition recognized the speaker on the basis of log likelihood. It recalculates the log likelihood of voice vector and compares it to the pre stored value of log likelihood. If it matches the entire log value from the database of specified PIN code then it provides access to the entire speaker to ATM.

5. Experimental Results

In this research, two phases are implemented to obtain experimental results.

- Registration phase
- Recognition phase

In first phase, the static voice sample is used to extract and trained the features and finally stored to the Speaker's database. It stores the entire trained features against the PIN code of specified speaker with his name. Entire features are extracted in 15 MFCC delta coefficients. That represents the pitch of human voice in the form of frequency on Mel scale. Delta coefficients are calculated to these Mel Coefficients and then trained using HMM forward backward algorithm. It results (Figure 4) in the log likelihood of entire voice and used to store in Speaker's database.

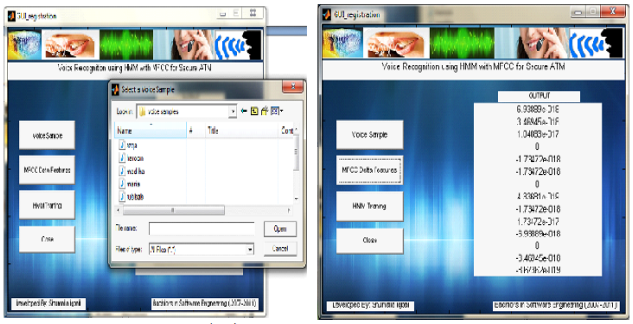


Fig. 4 Registration phase

Recognition phase works in real time scenario on application ATM. It resulted (Figure 5) to access the user account after verifying the PIN code as well as the log likelihood to the pre stored value against the entire speaker in Speaker's database.

Fig. 5 Recognition phase (ATM)

The results are tested against the specified objectives of proposed system. The developed system is tested by taking 3 speech samples from each speaker with the sampling frequency of 8 KHz. Voice features were extracted from

30 ms frame duration and 20 ms overlapping with the previous frame. The speech sample of 2 sec with noise is used to extract features and then trained using HMM (forward-backward) algorithm. The proposed MFCC features are used to expect the high accuracy in extracting the vocal features of voice. The HMM algorithm is anticipate to get best results in identification system. Accuracy rate shows the percentage of correctly identified test samples by the system. It is obtained by

Accuracy= number of correctly identified test samples/ total number of test samples = 26/ 30*100= 86.67%

Proposed system shows the accuracy of 86.67% as here total 30 test samples of each 10 speakers are being used to identify, where 26 out of total test samples are being correctly identified and correctly rejected with noise factor.

Table 1: Speaker's verification results

Number of speakers	Correct acceptance (with noise)	Correct rejection (with noise)	False acceptance (with noise)	False rejection (with noise)
Speaker 1				
Sample 1	√			
Sample 2	√			
Sample 3	√			
Speaker 2				
Sample 1	√			
Sample 2		√		
Sample 3	√			
Speaker 3				
Sample 1		√		
Sample 2	√			
Sample 3			√	
Speaker 4				
Sample 1		√		
Sample 2			√	
Sample 3	√			
Speaker 5				
Sample 1	√			
Sample 2	√			
Sample 3	√			
Speaker 6				
Sample 1				√
Sample 2	√			
Sample 3		√		
Speaker 7				
Sample 1		√		
Sample 2	√			
Sample 3			√	
Speaker 8				
Sample 1	√			
Sample 2	√			
Sample 3		√		
Speaker 9				
Sample 1	√			
Sample 2		√		
Sample 3		√		
Speaker 10				
Sample 1	√			
Sample 2	√			
Sample 3	√			

The error rate is calculated by

Total error of verification system= false accepted+ false rejected =3+1=4/30*100=13.33%

The error rate of total test samples that are being false rejected or accepted is 13.33%

The testing phase shows the efficiency of proposed system up to 86.67% with the error rate of 13.33% as depicted by the test results given in Table 1.

6. Conclusion

It is concluded that the proposed research uses the technique of MFCC to extract unique and reliable human voice feature pitch in the form of Mel frequency and trained and recognized using HMM log likelihood methodology. It comprises two security measures, PIN code as well as voice features to give more security to the ATM application. It represents the best efficiency up to 86.67% with the error rate of 13.33% on the basis of 30 test samples of 10 speakers (3 test samples per speaker).

The future work to leading this system is to provide the secure transmission of voice database to other branches of entire bank. Introducing encryption and decryption to transmit voice database and providing the facility to user to access his ATM account from any branch of bank. By taking the physiological features with behavioral features for recognition of person with voice as well as with his physical movement of mouth to make ATM more secure can tends to be a new research work.

References

[1] Debnath Bhattacharyya, Rahul Ranjan, Farkhod Alisherov A. and Minkyu Choi, "Biometric Authentication: A Review", International Journal of u- and e- Service, Science and Technology Vol. 2, No. 3, September, 2009

[2] Judith A. Markowitz, "Voice Biometrics", September 2000/Vol. 43, No. 9 Communications of the ACM

[3] http://en.wikipedia.org/wiki/Speaker_recognition

[4]http://www.globalsecurity.org/security/systems/biometrics-voice.htm

[5]Shi-Huang Chen and Yu-Ren Luo," Speaker Verification Using MFCC and Support Vector Machine", Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol I, IMECS 2009, March 2009

[6]Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques" Journal Of Computing, Volume 2, Issue 3, March 2010

[7]Ibrahim Patel and Dr. Y. Srinivas Rao, "Speech Recognition Using Hmm With Mfcc- An Analysis Using Frequency Specral Decomposion Technique", an International Journal (SIPIJ) Vol.1, No.2, December 2010

[8] Anjali Bala, Abhijeet Kumar and Nidhika Birla, "Voice Command Recognition System Based On MFCC and DTW"

Anjali Bala et al. / International Journal of Engineering Science and Technology Vol. 2 (12), 2010, 7335-7342
[9]http://en.wikipedia.org/wiki/Mel_scale
[10] Lawrence R. Rabiner, Fellow, IEEE 'A Tutorial On Hidden Markov Model And Selected Applications In Speech Recognition, Proceedings Of The IEEE, Vol. 77, No. 2, February 1989
[11]http://www.cs.brown.edu/research/ai/dynamics/tutorial/Documents/HiddenMarkovModels.html
[12]http://algoservice.com/qt/data/used-algoriyhms-reference.pdf
[13] Ms. G. Nathiya, Mrs. S. C. Punitha and Dr. M. Punithavalli," An Analytical Study on Behavior of Clusters Using K Means, EM and K*Means Algorithm", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, March 2010

Shumaila Iqbal She is a graduate student of Software engineering from Fatima Jinnah Women University. She Participated in Quiz Competition held by Sidra Tabassum, Chair IEEE Student Chapter 09.

Tahira Mahboob
She is a registered Engineer with the Pakistan Engineering Council. She received her bachelors degree from University of Engineering & Technology, Lahore in 2007. Currently enrolled in the MS/MPhil Computer Engineering program at Center for Advanced Studies in Engineering CASE(2010), UET Taxila. She has industry experience in telecom sector (Intelligent networks & VAS) Currently serving as a lecturer at Fatima Jinnah Women University, Rawalpindi. She has supervised thesis/projects at Bachelors and Masters Degree Programs in mobile/computer communications, voice recognition, mobile automation and cloud computing. Her area of interests are computer/mobile communications & networks, information security, mobile automation and adhoc/ sensor networks.

Malik Sikandar Hayat Khiyal
Dr. **Malik Sikandar Hayat Khiyal** is Chairperson Department of Computer Sciences and Software Engineering at Fatima Jinnah Women University, Pakistan. He received his M.Sc degree from Quaid-e-Azam University, Islamabad. He got first position in the faculty of Natural Science of the University. He was awarded the merit scholarship for Ph.D. He received his Ph.D. degree from UMIST, Manchester, U.K. He developed software of underground flow and advanced fluid dynamic techniques. His areas of interest are Numerical Analysis, Modeling and Simulation, Discrete structure, Data structure, Analysis of Algorithm, Theory of Automata and Theory of Computation. He has more than hundred research publications in National and International Journals and Conference proceedings.

Information Extraction and Webpage Understanding

M.Sharmila Begum[1], L.Dinesh[2] and P.Aruna[3]

[1] Assistant professor, Department of Software Engineering, Periyar Maniammai University Thanjavur,Tamil Nadu,India

[2] Department of Information Technology, Periyar Maniammai University Thanjavur,Tamil Nadu,India

[3] Assistant professor ,Department of Software Engineering, Periyar Maniammai University Thanjavur,Tamil Nadu,India

Abstract

The two most important tasks in information extraction from the Web are webpage structure understanding and natural language sentences processing. However, little work has been done toward an integrated statistical model for understanding webpage structures and processing natural language sentences within the HTML elements. Our recent work on webpage understanding introduces a joint model of Hierarchical Conditional Random Fields (HCRFs) and extended Semi-Markov Conditional Random Fields (Semi-CRFs) to leverage the page structure understanding results in free text segmentation and labeling. In this top-down integration model, the decision of the HCRF model could guide the decision making of the Semi-CRF model. However, the drawback of the topdown integration strategy is also apparent, i.e., the decision of the Semi-CRF model could not be used by the HCRF model to guide its decision making. This paper proposed a novel framework called WebNLP, which enables bidirectional integration of page structure understanding and text understanding in an iterative manner. We have applied the proposed framework to local business entity extraction and Chinese person and organization name extraction. Experiments show that the WebNLP framework achieved significantly better performance than existing methods.

Keywords: *Natural language processing, Webpage understanding, Information Extraction, Conditional Random Fields*

1. Introduction

The World Wide Web contains huge amounts of data. However, we cannot benefit very much from the large Amount of raw web pages unless the information within them is extracted accurately and organized well. Therefore, information extraction plays an important Role in Web knowledge discovery and management. Among various information extraction tasks, extracting Structured Web information about real-world entities (such as people, organizations, locations, publications, products) Has received much attention of late. However, little work has been done toward an integrated Statistical model for understanding web page structures and processing natural language sentences within the HTML Elements of the web page. Our recent work on Web object Extraction has introduced a template in dependent approach to understand the visually out structure of a webpage and to effectively label the HTML elements with attribute names of an entity. Our latest work on web page understanding introduces a joint model of The Hierarchical Conditional Random Fields (HCRFs) model and the extended Semi-Markov Conditional Random Fields (Semi-CRF's) model to leverage The page structure understanding results in free text Segmentation and labeling. The HCRF model can reflect the structure and the Semi CRF model can make use of the gazetteers. In this top down integration model, the decision Of the HCRF model could guide the decision of the Semi CRF model. However, the drawback of the top-down Strategy is that the decision of the Semi-CRF model could not be used by the HCRF model to refine its decision making. In this paper, we introduce a novel frame work called WebNLP at enables bidirectional integration of page structure understanding and text understanding in an iterative manner. In this manner, the results of page structure understanding and text understanding can be used to guide the decision making of each other, and the

performance of the two understanding procedures is boosted iteratively.

1.1 Overview

The World Wide Web contains huge amounts of data. However, we cannot benefit very much from the large amount of raw web pages unless the information within them is extracted accurately and organized well.Webpage understanding introduces a joint model of Hierarchical Conditional Random Fields (HCRFs) and extended Semi-Markov Conditional Random Fields (Semi-CRFs) to leverage the page structure understanding results in free text segmentation and labeling. In this top-down integration model, the decision of the HCRF model could guide the decision making of the Semi-CRF model. However, the drawback of the top down integration strategy is also apparent, i.e., the decision of the Semi-CRF model could not be used by the HCRF model to guide its decision making.
Our recent work on Web object extraction has introduced a template-independent approach to understand the visual layout structure of a webpage and to effectively label the HTML elements with attribute names of an entity.

In this paper, we introduced the Web NLP framework for webpage understanding. It enables bidirectional integration of page structure understanding and natural language understanding.Specifically, the Web NLP framework is composed of two models, i.e., the extended HCRF model for structure understanding and the extended Semi-CRF model for text understanding. The performance of both models can be boosted in the iterative optimization procedure.The experimental results show that the Web NLP framework performs significantly better than the state-of the- art algorithms on English local entity extraction and Chinese named entity extraction on WebPages.

2. Literature Survey

2.1 Information Extraction

IE technology has not yet reached the market but it could be of great significance to information end-user industries of all kinds, especially finance companies, banks, publishers and governments. For instance, finance companies want to know facts of the following sort and on a large scale: what company take-overs happened in a given time span; they want widely scattered text information reduced to a simple data base. Lloyds of London need to know of daily ship sinkings throughout the world and pay large numbers of people to locate them in newspapers in a wide range of languages. All these are potential uses for IE.

2.2 Empirical Methods in Information Extraction

The first large-scale, head-to-head evaluations of NLP systems on the same text-understanding tasks were the Defense Advanced Research Projects Agency–sponsored Message-Understanding Conference (MUC) performance evaluations of information-extraction systems. Prior to each evaluation, all participating sites receive a corpus of texts from a predefined domain as well as the corresponding answer keys to use for system development. The answer keys are manually encoded templates—much like that capture all information from the corresponding source text that is relevant to the domain, as specified in a set of written guidelines. After a short development phase, the NLP systems are evaluated by comparing the summaries each produces with the summaries generated by human experts for the same test set of previously unseen texts. The comparison is performed using an automated scoring program that rates each system according to measures of recall and precision.

2.3 Extracting Structured Data from Web Page

The World Wide Web is a vast and rapidly growing source of information. Most of this information is in the form of unstructured text, making the information hard to query. There are, however, many web sites that have large collections of pages containing structured data, *i.e.*, data having a structure or a *schema*. These pages are typically generated dynamically from an underlying structured source like a relational database. An example of such a collection is the set of book pages in Amazon. The data in each book page has the same schema, *i.e.*, each page contains the title, list of authors, and price of a book and so on.

2.4 Wrapper Induction Efficiency &Expressiveness

Wrapper is a procedure to extract all kinds of data from a specific web source. First find a vector of strings to delimit the extracted text.

Motivations: hand-coded wrapper is tedious and error-prone. How about web pages get changed? Wrapper induction –- automatically generate wrapper is a typical machine earning technology. Actually we are trying to learn a vector of delimiters, which is used to instantiate some wrapper classes (templates), which describe the document structure free text & Web pages. A good wrapper induction system should be:

Expressiveness: concern how the wrapper handles a particular web site.

Efficiency: how many samples are needed? How much computational is required?

2.5 Wrapper Maintenance Machine Learning Approach

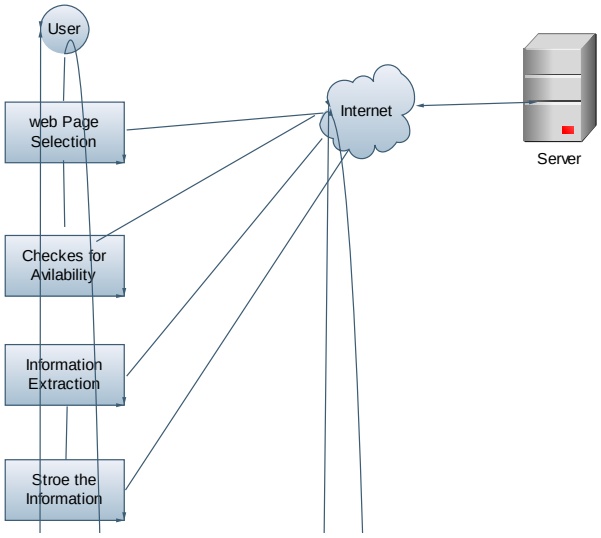
A Web wrapper is a piece of software that enables a Web source to be queried as if it were a database. The types of sources that this applies to are what are called semi structured sources. These are sources have no explicit structure or schema, but have an implicit underlying structure. Even text sources such as email messages have some structure in the heading that can be exploited to extract the date, sender, addressee, title, and body of the messages. Other sources, such as an online catalog, have a very regular structure that can be exploited to extract all the data automatically.

2.6 Hierarchical Wrapper Induction for Semi structured Information Sources

Web pages are intended to be human readable, there are some common conventions for structuring HTML documents. For instance, the information on a page often exhibits some hierarchical structure; furthermore, semi structured information is often presented in the form of lists of tupelos, with explicit separators used to distinguish the different elements. With these observations in mind, we developed the embedded catalog (EC) formalism, which can describe the structure of a wide-range of semi structured documents.

3. Implementation

Extracting information includes these modules to extract the structured data and natural language sentences with in the HTML elements of the webpage.



- ◊ Admin
- ◊ Page Reader
- ◊ Information Extraction
- ◊ Security
- ◊ Information maintenance

3.1 Admin

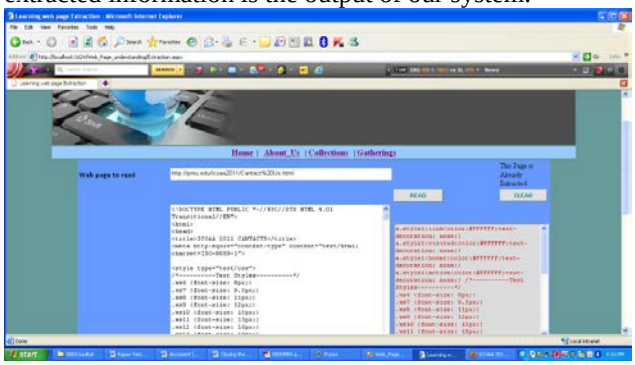
This module provides the facility to control all the operations done by our system. This module completely gives the rights to a single person. This module facilitate the update operation of the data secured by our systemAdmin module is commander module of our system.This module is the central module which integrates other modules. Admin module gives rights to a single person to perform the information extraction.Admin can update the database.Admin only can delete the records from database.

3.2 Page Reader

This module is the heart of our system.This module reads and understands the web pages and its structure.Two algorithms are used by this module.This module provides the facility to read the web page in both directions.Page reader module Reads the enter page source from the server.This module facilitates the source extraction from the server.This produces the output as collection of text information and tags,And also this produces the image links and other links provided in the current site.

3.3 Information Extraction

This is one of the major (Heart) modules of this system.This facilitates the easy ways for information extraction by our system.This module extracts information from the page reader.This module facilitates our system to extract the pure text information from the read source.Then this module extracts the fields and value from the source.This module only extracts exact information.The extracted information is the output of our system.

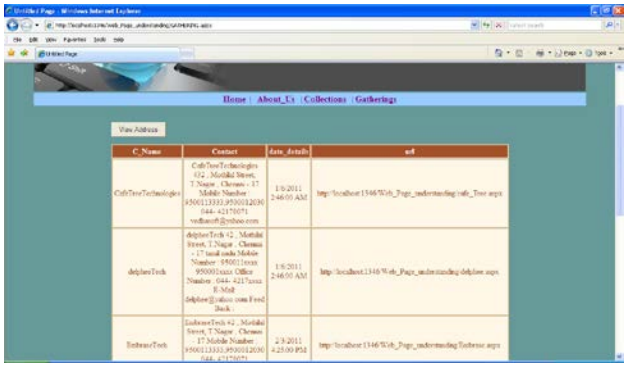


3.4 Security

This module provides the secure operations on our systems.This module allows the secure access by authorized persons only. This module checks for the security verifications like copyrights.This module facilitates our system while understanding the structure of the given input website.This module also restricts from unauthorized access of our application.

3.5 Information maintenance

This is another module of our system.This module provides the facility to store and maintaining the information.It facilitates to store details like Extracted sites and the information extracted from the sites.Information maintenance module facilitates our system for maintaining extracted information by our system.This module provides the facility to store the information. This is also provides the facility to retrieve the information from database



4. Conclusions

Webpage understanding plays an important role in Web search and mining. It contains two main asks, i.e., page structure understanding and natural language understanding. However, little work has been done toward an integrated statistical model for understanding webpage structures and processing natural language sentences within the HTML elements.

In our system, we introduced the WebNLP framework for webpage understanding. It enables bidirectional integration of page structure understanding and natural language understanding. Specifically, the WebNLP framework is composed of two models, i.e., the extended HCRF model for structure understanding and the extended Semi-CRF model for text understanding. The performance of both models can be boosted in the iterative optimization procedure. The auxiliary corpus is introduced to train the statistical language features in the extended Semi-CRF model for text understanding, and the multiple occurrence features are also used in the extended Semi-CRF model by adding the decision of the model in last iteration. Therefore, the extended Semi-CRF model is improved by using both the label of the vision nodes assigned by the HCRF model and the text segmentation and labeling results, given by the extended Semi-CRF model itself in last iteration as additional input parameters in some feature functions; the extended HCRF model benefits from the extended Semi-CRF model via using the segmentation and labeling results of the text strings explicitly in the feature functions. The WebNLP framework closes the loop in webpage understanding for the first time. The experimental results show that the WebNLP framework performs significantly better than the state-of the-art algorithms on English local entity extraction and Chinese named entity extraction on WebPages.

5. References

[1] J. Zhu, Z. Nie, J.-R. Wen, B. Zhang, and W.-Y. Ma, “Simultaneous Record Detection and Attribute Labeling in Web Data Extraction,” Proc. Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp 494-503, 2006.

[2] Z. Nie, Y. Ma, S. Shi, J.-R. Wen, and W.-Y. Ma, “Web Object Retrieval,” Proc. Conf. World Wide Web (WWW), pp. 81-90, 2007.

[3] J. Zhu, B. Zhang, Z. Nie, J.-R. Wen, and H.-W. Hon, "Webpage Understanding: An Integrated Approach," Proc. Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 903-912, 2007

[4] Y. Zhai and B. Liu, "Structured Data Extraction from the Web Based on Partial Tree Alignment," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 12, pp. 1614-1628, Dec. 2006.

[5] D. Downey, M. Broadhead, and O. Etzioni, "Locating Complex Named Entities in Web Text," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI), pp. 2733-2739, 2007.

[6] O. Etzioni, M.J. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D.S. Weld, and A. Yates, "Unsupervised Named- Entity Extraction from the Web: An Experimental Study," Artificial Intelligence, vol. 165, no. 1, pp. 91-134, 2005.

[7] Y. Zhai and B. Liu, "Structured Data Extraction from the Web Based on Partial Tree Alignment,"
IEEE Trans. Knowledge and Data Eng., vol. 18, no. 12, pp. 1614-1628, Dec. 2006.

[8] R. Song, H. Liu, J.-R. Wen, and W.-Y. Ma, "Learning Block Importance Models for Web Pages,"Proc. Conf. World Wide Web (WWW), pp. 203-211, 2004.

[9] J.D. Lafferty, A. McCallum, and F.C.N. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," Proc. Int'l Conf. Machine Learning (ICML), pp. 282-289, 2001.

[10] A. Chen, F. Peng, R. Shan, and G. Sun, "Chinese Named Entity Recognition with Conditional Probabilistic Models," Proc. Fifth SIGHAN Workshop Chinese Language Processing, pp. 173-176, 2006.

[11] D. DiPasquo, "Using HTML Formatting to Aid in Natural Language Processing on the World WideWeb,"http://citeseer.ist.psu.edu/dipasquo98using.html, 1998.

[12] C. Jacquemin and C. Bush, "Combining Lexical and Formatting Cues for Named Entity Acquisition from the Web," Proc. 2000 Joint SIGDAT Conf. Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 18

Sharmila Begum received M.E degree in Computer Science and Engineering. She is currently working as a Assistant Professor in Department of Software Engineering in Periyar Maniammai University Thanjavur Tamilnadu India. She has Presented several papers in international conferences and published few papers in PMU journal and published a book named Design and Analysis of Algorithms her research areas are Data Mining, Bio-Medical, OOAD, Networking and Web Programming.

Dinesh received M.Sc degree [5 Years Integrated] in Software Engineering.He is currently pursing his M.E Software Engineering in Periyar Maniammai University Thanjavur Tamilnadu India.

Aruna received MCA and M.Phil degree in Computer Application. She is currently working as a Assistant Professor in Department of Software Engineering in Periyar Maniammai University Thanjavur Tamilnadu India. She has presented several papres in International conferences and her research area is Mobile Adhoc Network.

Literature Survey on Design and Implementation of Processing Model for Polarity Identification on Textual Data of English Language

Aparna Trivedi, Apurva Srivastava, Ingita Singh, Karishma Singh and Suneet Kumar Gupta

Information Technology(B. Tech IV year), Gautam Buddh Technical University, ABES Engineering College Ghaziabad, 201009, India

Information Technology (B. Tech IV year), Gautam Buddh Technical University, ABES Engineering College Ghaziabad, 201009, India

Information Technology (B. Tech IV year), Gautam Buddh Technical University, ABES Engineering College Ghaziabad, 201009, India

Information Technology (B. Tech IV year), Gautam Buddh Technical University, ABES Engineering College Ghaziabad, 201009, India

Information Technology (Associate Professor), Gautam Buddh Technical University, ABES Engineering College Ghaziabad, 201009, India

Abstract

This literature work is a survey about Sentiment Analysis of textual data of English language and some of its previous works. Basically sentiment analysis identifies the view point or opinion of a text. For example, classifying a movie review as "Thumbs up" or "Thumbs down".

Several public opinion surveys from multiple polling organization and people's aggregate opinion on a topic can be assessed. This survey includes previous works which show how this technique has evolved over the past one and a half decade expanding its horizon and reaching out to almost all areas such as reviews of products, movies etc., travel advice, stock market predictions and in other decision making areas.

Keywords: *Opinion Mining, Sentiment Analysis, Polarity Identification.*

1. Introduction

Natural Language Processing is a domain of computer science and scientific study of human language i.e. linguistics which is related with the interaction or interface between the human (natural) language and computer. Basically NLP commenced as a sub-field of artificial intelligence. Opinion mining or Sentiment analysis refers to a broad area of Natural Language Processing and text mining. It is concern not with the topic a document is about but with opinion it expresses hat is the aim is to determine the attitude (feeling, emotion and subjectivities) of a speaker or writer with respect to some topic to determine opinion polarity. Initially it was applied for classifying a movie as good or bad based on positive or negative opinion. Later it expanded to star rating

predictions, product reviews travel advice and other decision making processes.

According to the survey performed by Bo Pang and Lillian Lee, Sentiment analysis identifies the view points of a text. For example, classifying a movie review as thumbs up (recommended) or thumbs down (not recommended). Previous methods focused on selective lexical features (e.g. word "Good"), then classifying document according to the number of such features that occur anywhere within it. But in contrast later following process were followed:

- Identify the sentences in the given input text as subjective or objective.
- Select and apply a standard machine learning classifier to the extracted result.

This could prevent the polarity classifier from considering misleading, ambiguous or irrelevant text. For example, the sentence "The protagonist tries to protect her good name" holds the word "Good", but it reports nothing about author's opinion and could also be implanted in a negative way.

Our work is based on this technique of Sentiment analysis using polarity classification of textual data. In this, we estimate the percentage of positivity or negativity of input text by first tagging all the adjectives, adverbs using a POS (Part of Speech) tagger (Marks words in the input text corresponding to a particular part of speech). Then we estimate the positivity or the negativity of the extracted adjectives corresponding to its value in the SentiWordNet (derived from WordNet, a lexical database, where numerical value indicating polarity sentiment, i.e. positive or negative, information corresponds to each word in it). In order to estimate sentiment orientation we count the positive and negative terms values. Finally, we assign estimated polarity to the given corpus.

2. Evolution

According to the paper titled "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews" by Peter D. Turney, presented in the Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (2002), in Philadelphia, Pennsylvania, a major application of sentiment analysis of textual data is in the classification of any review. Example: The semantic orientation of phrases with the help of simple unsupervised learning algorithm. The whole process of classification can be summaries into simple three steps:

1. Identify phrases in the given corpus containing adjectives or adverbs (using a part-of-speech tagger given by Brill in 1994).
2. Approximate the semantic orientation of the identified phrase.
3. Based on the sentiment orientation classify the given input text.

This algorithm makes use of PMI-IR to calculate semantic orientation. Peter D. Turney experimented with 410 reviews of various domains and concluded that the algorithm accomplishes an average accuracy of about 74%. But for movie review its about 66% while 82%-84% for automobiles and banks. The limitations identifier in this work of Peter D. Turney was the time needed for queries which can be eliminated by development in hardware. The level of accuracy can be improved by gelling semantic features with some distinct features of a supervised classification algorithm. This work has its nearness to the work of "Predicting the semantic orientation of adjectives" by Hatzivassiloglow and Mc Keown presented at the Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL in 1997. The use of four step algorithm which:

1. Removes conjunction to isolate adjectives.
2. Uses a supervised learning algorithm to label adjectives into groups of same or different semantic orientation and result in the graph where nodes denote adjectives and links denote similarity or difference in semantic orientation.
3. Using a clustering algorithm, the graph is processed to give two subsets of adjectives: Positive and negative.
4. If the frequency of positive adjective is high, then the text is positive else negative.

But the algorithm overall is complex and improved in various fields.

Another work in this field was R.M Tong's "An operational system for detecting and tracking opinions in on-line discussions" at the ACM SIGIR 2001 Workshop on Operational Text Classification in 2001. The system trails online discussion and gives a graph for positive and negative sentiments looks for phrases like "bad acting", "awesome music", "uneven editing" etc. In This edition of phrases to a special lexicon as tagging of sentiment as positive or negative is done manually. But this work was specific for movies.

Further in this field, a research work on "Sentiment classification of reviews using SentiWordNet" was conducted by Bruno Ohana and Brendan Tierney, of Dublin Institute Of Technology, and presented in the 9th I.T & T Conference, 2009. They used automatic methods for speculating the course of subjective content on textual data. SentiWordNet (opinion lexicon) is basically used to classify automatic sentiment of film reviews and the research done elaborates the results produced. The research goes one step ahead through extending the use of SentiWordNet by building the set of significant features and applying to the machine learning classifier. The set of relevant features provided substantial enhancement over baseline term counting methods. The important conclusion drawn indicated that SentiWordNet has now emerged as

an indispensable tool for sentiment classification tasks and further progress can be made in its user and its usage along with other techniques. The research associates words and their synonyms present in Synsets with two numerical number ranging from 0 to 1, each denote the SentiWordNet's positive and negative bias.

SentiWordNet's one of the prominent features is that, in this a term can posses can both positive and negative score to have non-zero values. Stanford part of speech tagger was used in the research to correctly associate scores to terms and then scores for each term was found. The ratio between scores and number of terms was found and overall score was calculated. The document was divided into sections and scoring was performed section-wise and in the end the final sentiment of polarity was analysed.

"From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series", was proposed in May 2010, in which Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge and Noah A. Smith analysed several public opinion surveys from multiple polling organisations on consumer confidence and political over 2008 to 2009 period and found that they co-relate to sentiment word frequencies in contemporary twitter messages. The results vary across data sets and sometimes correlations were as high as 80%. For example, if we want to know the extent to which U.S. population likes or dislikes Obama, polling methodology is done. It was extensively developed through 20th century (Krosnick, Judd and Wittenbrink 2005). From text, population's aggregate opinion on a topic can be assessed and then the task can be broken down into two sub problems:

1. Message Retrieval – When we identify the messages relating to topic.
2. Opinion Estimation – Determine whether messages express a positive or negative opinion or just news about the topic.

Tamara Martin-Wanton and Aurora Pons-Porrata gave a paper "Opinion polarity detection" in which an unsupervised algorithm was used for polarity of opinion which uses a word sense disambiguation algorithm to determine the correct sense of the word in the opinion. This proposed method does not depend on the knowledge domain and can be extended to other languages. The resources used by the author for this method is:

1. WordNet (Lexical database).
2. SentiWordNet (Lexical Resource).
3. A subset of General Inquirer (English Dictionary).

The two basic components of this method are:

1. Word sense disambiguation
2. Determination of polarity

Word sense disambiguation identifies the correct senses of the terms and in the determination of the polarity we determine the polarity of the opinion. The uniqueness of this method is using standard external resources along with word sense disambiguation for determining polarity of the opinions. Thus, this method is independent of knowledge domain and can be extended to other languages. Because of wrong annotations of SentiWordNet, there may be failure of method in many cases.

One of the most recent paper is "The Truth About Sentiment and Natural Language Processing" published by Synthesio (a global, multilingual, Social Media Monitoring and Research company) in March'11 which focuses on how brands can ascertain what opinions people have about their brand through sources like social media, blogs, online newspaper and magazines etc. Even if the sentiment analysis is inappropriate and no social media assistance could prove that this technology could accurately access sentiment on a precise topic, but by stalking and leaning it over time we can examine the pattern for changes since we are presumptuous that the in correctness will be constant over time. However there is no proof of this yet.

In 2009 and 2010, Amitava Das and Sivaji Bandyopadhyaya of Jadavpur University presented their research paper "Phrase-level Polarity Identification for Bangla" and "SentiWordNet for Indian Languages", respectively, which emphasize on opinion polarity classification on news texts using Support Vector Machine (SVM) for popular Indian native language Bengali. The contemporary system present directs the course of an opinionated phrase to positive or negative. A pre-requisite for identifying the direction of opinion requires the categorization of texts into subjective or objective. The reason being objective text cannot be predicted by definition. The system uses a combined approach which co-ordinates well with lexicon entities and linguistic syntactic features and the classifier used in the research is rule based subjectivity. The results have accuracy of 70.04% and a recall of 63.02%. The limitation of the research lies on the usage of log-linear functions models like SVM. The major drawback being that a well distinct decision boundary cannot be formed from the conjunction of provided features. The disadvantage can be overcome by providing the conjunctions explicitly as an integrated unit of feature vector, by stating the features as a classical word lattice model. Finally, the post processor gives to the chunk head a polarity value which will be directly proportional to the chunk head's resultant polarity domain.

Now, presently research is done in the advancement of present system in the way of progressive methods for formation of opinions based on their polarity class.

3. Conclusion

This paper represents a lexical unison based measurement of sentiment intensity and polarity in text and its application in various fields. We are further working on how best to analyze the psychological effect of text by exploiting available resources and evaluating polarity of the text.

4. References

1. Tong, R.M. 2001. An operational system for detecting and tracking opinions in on-line discussions. *Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification* (pp. 1-6). New York, NY: ACM.

2. Turney, P.D. 2001. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. *Proceedings of the Twelfth European Conference on Machine Learning* (pp. 491-502). Berlin: Springer-Verlag.

3. Hatzivassiloglou, V., & McKeown, K.R. 1997. Predicting the semantic orientation of adjectives. *Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL* (pp. 174-181). New Brunswick, NJ: ACL.

4. Pang, B., and Lee, L. 2008. Opinion Mining and Sentiment Analysis. Now Publishers Inc.

5. Krosnick, J. A.; Judd, C. M.; and Wittenbrink, B. 2005. The measurement of attitudes. The Handbook of Attitudes 2176.

6. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series : Brendan O'Connory, Ramnath Balasubramanyany, Bryan R. Routledgex, Noah A. Smithy.

7. Amitava Das and Sivaji Bandyopadhyay. Theme Detection an Exploration of Opinion Subjectivity. In Proceeding of Affective Computing & Intelligent Interaction (ACII 2009).

8. Peter Turney, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceeding of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics.

9. Sentiment Classification of Reviews Using SentiWordNet: Bruno Ohana , Brendan Tierney.

10. Synthesio- The Truth About Natural Language Processing, March 2011.

11. Agirre, E., Soroa, A., (2007). Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, 7-12.

12. OPINION POLARITY DETECTION: *Using Word Sense Disambiguation to Determine the Polarity of Opinions*- Tamara Martín-Wanton, Aurora Pons-Porrata, Andrés Montoyo-Guijarro, Alexandra Balahur.

13. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *EMNLP*, pages 79.86

Aparna Trivedi, currently pursuing B.Tech. from ABES Engineering College, Ghaziabad and involved in research work of Natural language Processing and Information Retrieval**.**

Apurva Srivastava, currently pursuing B.Tech. from ABES Engineering College, Ghaziabad and involved in research work of Natural language Processing and Information Retrieval**.**

Ingita Singh, currently pursuing B.Tech. from ABES Engineering College, Ghaziabad and involved in research work of Natural language Processing and Information Retrieval**.**

Karishma Singh, currently pursuing B.Tech. from ABES Engineering College, Ghaziabad and involved in research work of Natural language Processing and Information Retrieval**.**

Suneet Kumar Gupta, currently working as Associate Professor at ABES Engineering College, Ghaziabad and has many years of experience in teaching and research. Currently working on Natural language Processing and Information Retrieval.

Data Mining in Sequential Pattern for Asynchronous Periodic Patterns

Thodeti Srikanth
Research Scholar, Ph.D. (Computer Science),
Dravidian University, Andhra Pradesh, INDIA

Abstract- Data mining is becoming an increasingly important tool to transform enormous data into useful information. Mining periodic patterns in temporal dataset plays an important role in data mining and knowledge discovery tasks. This paper presents, design and development of software for sequential pattern mining for asynchronous periodic patterns in temporal database. Comparative study of various algorithms on sequential pattern mining for asynchronous periodic patterns is also carried out by taking artificial and real life database of glossary shop. The proposed system will be based on optimization of Efficient Mining of Asynchronous Periodic Pattern Algorithm (EMAP), which will be implemented for efficient mining of asynchronous periodic patterns in large temporal database.

Keywords- Sequential patterns, Temporal dataset, Knowledge discovery, Asynchronous Periodic patterns,

I. INTRODUCTION

Pattern mining plays an important role in data mining tasks. Various patterns have been introduced for different applications, e.g., frequent item sets and sequential patterns for transaction databases, frequent episodes in event sequences, and frequent continuities for inter transaction association. Periodic patterns are recurring patterns that have temporal regularities in time-series databases. Periodic patterns exist in many kinds of data. For example, tides, planet trajectories, somite formation, daily traffic patterns, and power consumptions all present certain periodic patterns. There are many emerging applications, including stock market price movement, earthquake prediction, telecommunication network fault analysis, repeat detection in DNA sequences and occurrences of recurrent illnesses, etc. The discovery of patterns with periodicity has been studied in several works. For example, Ozden et al. proposed the mining of cyclic association rules that reoccur in every cycle of the time span of the temporal database. Han et al. considered imperfect periodic patterns that reoccur for at least minconf percent of the cycles. Berberidis et al. further proposed an approximate periodicity detection algorithm. However, these studies considered only synchronous periodic patterns and did not recognize the misaligned presence of patterns due to the intervention of random noise. For example, assume that a temporal database contains a periodic pattern, “burger and maggi,” on Friday nights, from January to March. However, in April, the business has a big promotion for beer every Saturday. Therefore, many customers would buy burger on Saturday instead of Friday because of this promotion[8].

In this case, it would be desirable if the pattern can still be recognized when the disturbance is within some reasonable threshold. Therefore, in , Yang et al. extended the idea to find asynchronous periodic patterns. Yang et al.’s asynchronous periodic pattern problem aims at mining the longest periodic subsequence which may contain a disturbance of length up to a certain threshold. Formally, a valid subsequence with respect to a pattern P in a sequence D is a set of non overlapping valid segments, where a valid segment has at least min rep contiguous matches of P and the distance between any two successive valid segments does not exceed a parameter max dis. A valid subsequence with the most overall repetitions of P is called its longest valid subsequence. However, this model has some problems. . First, this model only focused on mining periodic patterns in temporal sequences of events. However, in real-world applications, we may find multiple events at one time slot in terms of various intervals (e.g., hour, day, week, etc.) as discussed in previous works. We refer to such databases as sequences of event sets. . Second, this model only focused on mining the longest sequence of a pattern, which can only capture part of the system’s behavior. For example, in the case when two successive, non overlapped segments with a disturbance larger than max dis, only the larger segment will be reported[7].

To address these problems, K. Huang proposed a novel SMCA algorithm which requires no candidate pattern generation as compared to previous technique. Their algorithm allows the mining of all asynchronous periodic patterns, not only in a sequence of events, but also in a temporal dataset with multiple event sets. They also proposed a dynamic hash-based validation mechanism which discovers all asynchronous type periodic patterns in a single scan of temporal dataset. Their four phase approach uses a sequence of algorithms to mine singular pattern, multiple pattern, maximal complex pattern and finally asynchronous periodic patterns. Each of these algorithms uses output of the last executed algorithm as their input. The main limitation of their algorithm is that, it not only mines the maximal complex pattern but also its subsets (single event patterns and multiple events patterns) using depth first search enumeration approach, thus wasting a considerable amount of processing time for mining

subsets. For large datasets having *i-patterns* where *i* is too large, a large amount of processing time is waste for mining singular and multi events *1-patterns* that are subsets of *i-patterns*. To increase the efficiency of mining asynchronous periodic patterns on large datasets, we propose a novel efficient algorithm E-MAP. Our propose algorithm finds all maximal complex patterns in a single step algorithm using a single dataset scan without mining single event and multiple events patterns explicitly, while asynchronous periodic patterns are mined using the same depth first search enumeration process as described in . The single dataset scan and single step mining approach makes the E-MAP much faster and efficient as compared to previous technique SMCA. The other feature of E-MAP is that, it requires less storage space as compared to SMCA. To check the effectiveness of our E-MAP approach, we also provide detailed experimental results on real and artificial datasets. Our different experimental results suggest that mining asynchronous periodic patterns using E-MAP is more efficient as compared to SMCA.

Database mining is motivated by the decision support problem faced by most large retail organizations. Development of bar-code technology has made able retail organizations to collect and store massive amounts of sales data, referred to as the basket data. A record in such data typically consists of the transaction date and the items bought in the transaction [1]. Very often, data records also contain customer-id, particularly when the purchase has been made using a credit card or a frequent-buyer card. Catalog companies also collect such data using the orders they receive.

A sequence database consists of sequences of ordered elements or events, recorded with or without a concrete notion of time [1]. There are many applications involving sequence data. Typical examples include customer shopping sequences, Web click streams, biological sequences, sequences of events in science and engineering, and in natural and social developments.

Sequential Pattern Mining

Sequence Pattern Mining is the mining of frequently occurring ordered events or subsequences as patterns [9]. An example of sequential pattern is “Customers who buy a canon digital camera are likely to buy an HP color printer within a month”[1]. For retail data, sequential patterns are useful for shelf placement and promotions. Also telecommunications and other businesses may also use sequential patterns for targeted marketing, customer retention and many other tasks. Other areas in which sequential patterns can be applied include Web access pattern analysis, weather prediction, production processes, and network intrusion detection analysis. Most studies of sequential pattern mining concentrate on categorical patterns [6]. The sequential pattern mining problem was first introduced by Agrawal and Srikant in 1995[1] based on their study of customer purchase sequences, as follows: ”Given a set of sequences, where each sequence consist of a list of events(or element) and each event consists of set of items, and given a user specified minimum support threshold of min_sup, sequential pattern mining finds all the frequent subsequences, that is, the subsequences whose occurrence frequency in the set of sequences is no less than min_sup.”

Definition 1: Let $X = \{x_1, x_2, x_{3,..} x_n\}$ be a set of different items. An element *e*, denoted by $< x_1, x_{2, ...} >$, is a subset of items belonging to X which appear at the same time. A sequence *s*, denoted by $< e_1; e_2; \ldots ; e_m >$, is an ordered list of elements. A sequence database *Db* contains a set of sequences, and | *Db* | represents the number of sequences. in Db. A sequence $\alpha = < a_1; a_2; \ldots ; a_n >$ is a subsequence of another sequence $\beta =< b_1; b_2; \ldots ; b_m >$ if there exist a set of integers, $1 \leq i_1 \leq i_2 \leq i_n \leq m$, such that a_1 is a subset of b_{i1}; a_2 is a subset of b_{i2} ; . . . and *an* is a subset of b_{in} . The sequential pattern mining can be defined as “Given a sequence database *Db* and a user-defined minimum support min sup, find the complete set of subsequences whose occurrence frequencies ≥ min sup *| *Db* |” .

II. RELATED WORK

As mentioned, the sequential pattern mining with a static database and with an incremental database is two special cases of the progressive sequential pattern mining. In the following, we introduce the previous works on the static sequential pattern mining, the incremental sequential pattern mining, and the progressive sequential pattern mining. Previous researchers have developed various methods to find frequent sequential patterns with a static database.

The assumption of having a static database may not hold in many applications. The data in real world usually change on the fly. When we deal with an incremental database, it is not feasible to remine the whole sequential patterns every time when the database increases because the remaining process is costly. To handle the incremental database, Parthasarathy et al. presented the algorithm ISM [2] using a lattice framework to incrementally update the support of each sequential pattern in equivalent classes. Masseglia et al. derived the algorithm ISE [3] to join candidate sequential patterns in original database with the newly increasing database. Cheng et al. introduced algorithm IncSpan [4], which utilized a special data structure named sequential pattern tree to store the projection of database. However, the incremental mining algorithms can only handle the incremental parts of the database. Because of the limitation of data structures maintained in their algorithms, they can only create new candidates but cannot delete the obsolete data in a progressive database. The deletion of an item from the database results in the reconstruction of all candidate item sets, which induces incredible amount of computing.

III. PROBLEM STATEMENT

Design & Develop an Automatic Data Mining System for Asynchronous Periodic Patterns. The proposed system will be based on Modified version of E-MAP algorithm compatible for temporal dataset. The entries for temporal dataset will be taken from a real life database of a grocery shop.
Let E be a set of all events. An event set is a nonempty subset of E. A temporal dataset D is a set of records where each record is a tuple in the form < tid, X > for time instant tid and event set X.

Table 1: A temporal data with 4 events (A, B, C, D) and 7 time instances

Time instance	Event
1	A,C
2	A,B
3	A,C
4	A,B,D
5	D
6	C
7	C

III. FORMAT OF DATABASE

Table 2 : Database

Seq_no	Element	Time_stamp
1	Magi	1
1	Soup	2
1	Soup	4
2	Burger	2
2	Icetea	3
2	Magi	4
2	Icetea	5
3	Soup	1
3	Icetea	2
3	Magi	3
3	Icetea	5
4	Berger	3
4	Icetea	4
4	Burger	5
4	Magi	4
5	Soup	2
5	Burger	5

IV. IMPLEMENTATION

This system tries to find segment of subsequence as a valid pattern depending upon the minimum repetition and threshold value given to it. It provides the graphical user interface to input these values.

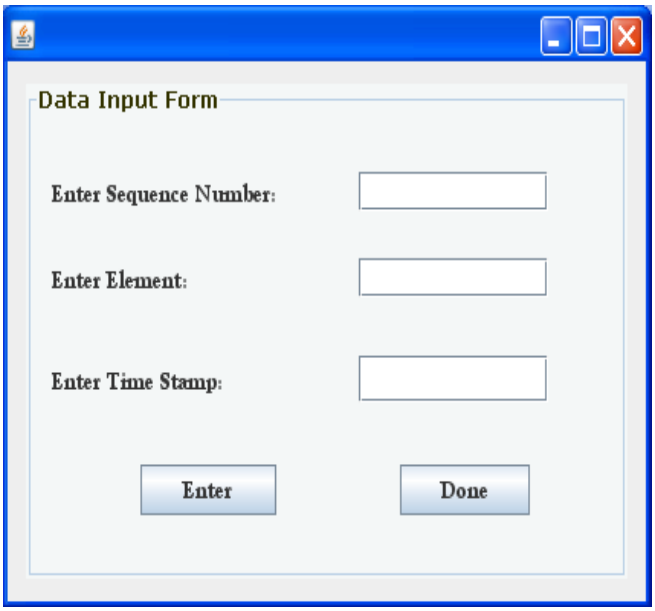


Fig 1 : GUI (Input data form)

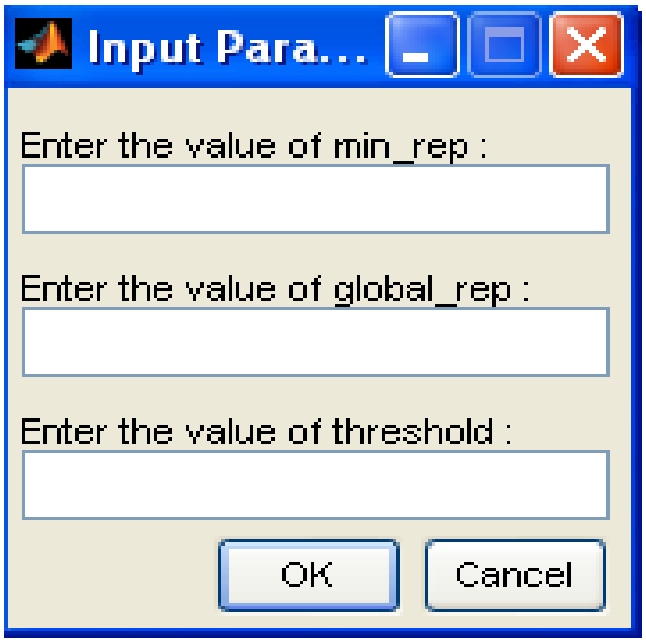


Fig 2 : GUI (Input parameters)

V. RESULTS & DISCUSSION

After implementation of the algorithm, we have obtained the result.
We tried to compare its performance with other mentioned algorithms on the basis of Space and time complexity .

Let $SP_{SMCA}(n)$: Space Complexity of SMCA
$SP_{EMAP}(n)$: Space Complexity of EMAP
$SP_{MEMAP}(n)$: Space Complexity of MEMAP

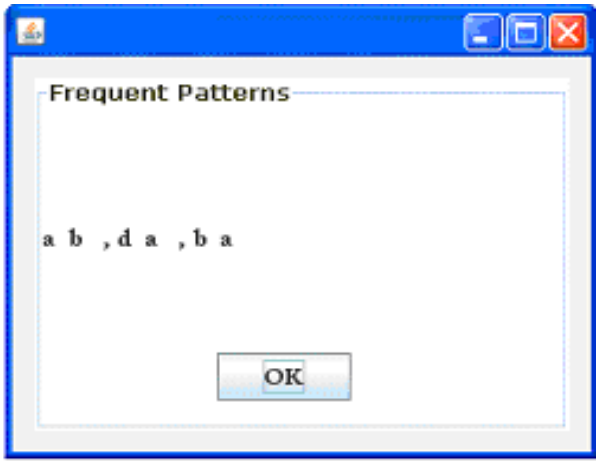


Fig 3 : Frequent patterns

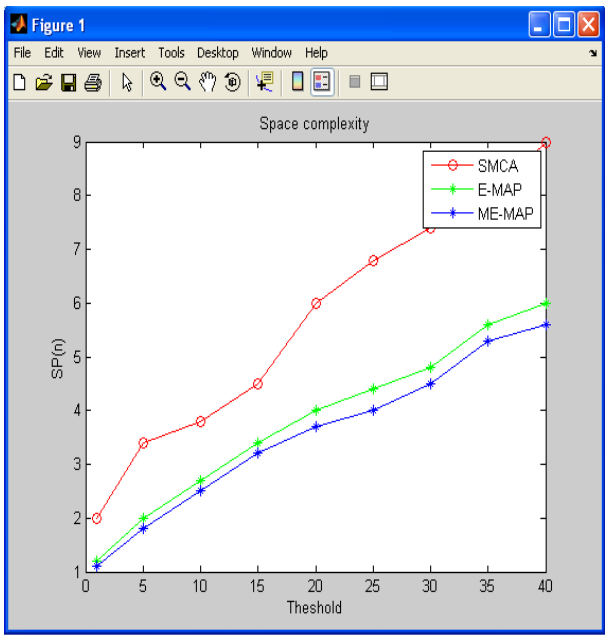


Fig 3 : Space Complexity

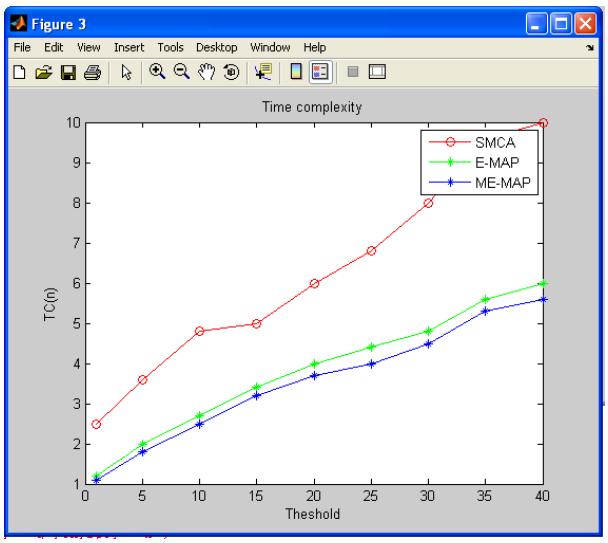


Fig 4 : Time Complexity

VI. CONCLUSION

In this paper we proposed a new algorithm ME-MAP for sequential pattern mining of asynchronous periodic patterns. We have studied the performances of SMCA, E-MAP and Modified E-MAP algorithms. The performance evaluation is done on the basis of time and space complexities of these algorithms. It is found that :

$SP_{SMCA}(n) \leq SP_{E\text{-}MAP}(n) \leq SP_{ME\text{-}MAP}(n)$

$TC_{SMCA}(n) \leq TC_{E\text{-}MAP}(n) \leq TC_{ME\text{-}MAP}(n)$

This shows the effectiveness of our approach.

VII. REFERENCES

[1] Agrawal, R. and Srikant, R. Mining sequential patterns. In Eleventh International Conference on Data Engineering, P. S. Yu and A. S. P. Chen, Eds. IEEE Computer Society Press, Taipei, Taiwan, pp. 3-14, 1995.

[2] S. Parthasarathy, M.J. Zaki, M. Ogihara, and S. Dwarkadas, "Incremental and Interactive Sequence Mining," Proc. 8th ACM Int'l Conf. Information and Knowledge Management (CIKM '99), pp. 251-258, 1999.

[3] F. Masseglia, P. Poncelet, and M. Teisseire, "Incremental Mining of Sequential Patterns in Large Databases," Data and Knowledge Eng., vol. 46, pp. 97-121, 2003.

[4] H. Cheng, X. Yan, and J. Han, "INCSPAN: Incremental Mining of Sequential Patterns in Large Database," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '04), pp. 527- 532, 2004.

[5] A. Balachandran, G.M. Voelker, P. Bahl, and P.V. Rangan, "Characterizing User Behavior and Network Performance in a Public Wireless LAN," Proc. ACM SIGMETRICS Int'l Conf. Measurement and Modeling of Computer Systems (SIGMETRICS '02), pp.195-205, June 2002.

[6] Jen-Wei Huang, Chi-Yao Tseng, Jian-Chih Ou, and Ming-Syan Chen. "A General Model for Sequential Pattern Mining with a Progressive Database", Knowledge And Data Engineering, vol. 20, no. 9, pp. 1153-1167, September 2008.

Thodeti Srikanth received his Master of Computer Applications degree from Kakatiya University, Andhra Pradesh, INDIA in 2004. He is pursuing Ph.D. (Computer Science) from Dravidian University, Andhra Pradesh, INDIA.

A Partitioning strategy for OODB

Dr. Sudesh Rani
Asstt. Prof.(Computer Science), Govt. College, Hisar
Kurukshetra University, Kurukshetra, Haryana, India

Abstract

An effective strategy for distributing data across multiple disks is crucial to achieving good performance in a parallel object-oriented database management system. During query processing, a large amount of data need to be processed and transferred among the processing nodes in the system. A good data placement strategy should be able to reduce the communication overheads, and, at the same time, to provide the opportunity for exploiting different types of parallelism in query processing, such as intra-operator parallelism, inter-operator parallelism, and inter-query parallelism. However, there exists a conflict between these two requirements. While minimizing interprocessor communication favors the assignment of the whole database to a small number of processors, achieving higher degree of parallelism favors the distributions of the database evenly among a large number of processors. A trade-off must be made to obtain a good policy for mapping the database to the processors.We need good heuristics to solve this and more complicated database allocation problems. In this paper, we propose some heuristics for partitioning an OODB so that the overall execution time can be reduced.

Keywords: Parallelism, Vertical partitioning, Horizontal partitioning, Query diameter.

1. Introduction

In order to achieve parallelism, the database needs to be partitioned over multiple components in a parallel system. For example, relations in Gamma ([4], [5]) are horizontally partitioned across all nodes with disk drives using one of four declustering strategies provided in the system: round-robin, hashed, range, and hybrid-range partitioned. However, none of the strategies is a clear winner in the performance analysis ([7], [8]). To decluster all relations across all nodes with disks is recognized as a serious mistake ([5]). A better solution used in Bubba ([3]) is to decluster a relation based on the \heat" (i.e., the cumulative access frequency) and the size of the relation. Since the ideal data placement changes continuously as the workload changes in time, Bubba repeatedly refines the data placement if the performance improvement is worth the work required to reorganize.

In a relational database environment, a relation may be accessed by several types of queries which require different sets of attributes. In order to improve the performance, attributes of the relation are divided into groups and the relation is projected into fragment relations according to these attribute groups. This process is called vertical partitioning. The fragments are assigned to different sites in distributed database systems to minimize the cost of accessing data by all queries.

There are trade-offs between horizontal and vertical partitioning methods. A general discussion of pros and cons on a decomposed storage system (DSM), which pairs each attribute value with the surrogate of its record, is reported by Copeland and Khoshafian ([2]). Several parallel database projects have employed some form of the same vertical data partitioning concept ([9], [10], [11], [14]). A simple file assignment problem, which deals with assigning files to different nodes of a computer network, has been studied extensively ([6]). However, most of the works assume that a request is made at one site and all the data for answering it is transferred to that site. This simple view of application cannot model the query processing strategy in a parallel database system. The simple file assignment problem is an NP-complete problem

As for the OODBs, it is recognized that object clustering is important to the performance ([1], [12], [13]). However, the clustering in OODBs is still an open research issue, and therefore the problem of declustering an OODB for a parallel system is a new challenge in research.

2. The Problem

The problem is to partition a given OODB and assign the partitions to the nodes in a multiprocessor system. It is assumed that the number of object classes in the database is larger than the number of processors in the system. Also, we assume that the processors are fully connected. This simplifies the problem so that we do not need to consider the effects of the network's physical topology. However, we can simulate different topologies by introducing various delays to different links.

We assume that the unit of distribution is class. In other words, classes are not allowed to be split, and each class must reside in one and only one node. Since we group all the data associated with an object class together, we can localize retrieval, manipulation, and user-defined operations and reduce the overall communication among processors. If we horizontally partition the classes and assign them to multiple processors, two sets of processors

need to communicate with each other when two classes want to exchange information. In addition, if a large number of processors work on the same class, this horizontal partition scheme does not provide a good environment for multiple queries to be executed in parallel when these queries access different classes. Thus, we choose class as the unit of partition in this study. However, if some classes are too large for one node to handle and we decide to split them, the heuristics presented in this paper can still be used to group the partial classes.

It is not easy to find a partition which is good for all the applications. A good partition for one application may not be suitable for another application. If we make a compromise for both applications, neither one will perform well. Therefore, we decide to partition the database based on the processing requirement of a single application which is characterized by a set of typical queries used in the application. By analyzing the query patterns in the set and the data characteristics of the database, we try to find a partition so that the execution time of the set of queries is minimized.

If we want to calculate the execution time of a query, an appropriate cost function is needed for modeling the parallel execution of the query. For a set of queries, the interaction and interference among queries will make it extremely difficult to formulate the cost function. Even if we can formulate the correct function, the problem of finding the minimum would be intractable. Therefore, instead of finding the best partition which gives the minimal execution time, we try to find some heuristic rules that will avoid bad partitions and give good performance.

3. Heuristics for Partitioning an OODB

The execution time of a query in a parallel environment consists of three components: CPU time, IO time, and communication time. Since the CPU time and IO time in each processor are the time the processor works on the query, we use the term "processing cost" to represent these two time components. It takes some communication time for a message to transfer from one processor to another. However, both the source and the destination processors can do other tasks during this time.

If the communication delay is short, one obvious bad solution for partitioning the database is to assign all object classes to one single node. In other words, we want to balance the processing load on the nodes as well as to reduce the communication cost among them. However, we cannot use the sum of the processing cost and the communication cost as the total cost for a partition because it is difficult to give a meaning to the combined cost. Also, if we use the combined cost to partition the database, we run the risk of having two equal cost partitions in which one has high processing cost and the other one has high communication cost. On the other hand, **if the communication delay is long,** we want to group classes that exchange large amount of data and reduce the length of the "path" through which messages and data must be transferred. Therefore, we try to find a combined heuristics for partitioning the database.

The heuristic method is based on the overall processing cost of each class referenced in a query. We measure the overall CPU time and IO time used for processing a class to represent the processing cost of that class in the query. When we consider the set of queries, we take the sum of the processing costs for the same class in all queries to represent the total processing cost for that class.

Figure 1 shows the example university database with a class number and a class size in parentheses (i.e., the number of instances) attached to each class. A set of 10 queries as shown in Figure 2 represents the processing requirement of a specific application that we want to partition the database for. In this example, we have 5 simple queries (queries 0, 1, 2, 3, and 4) and 5 complex queries (queries 5, 6, 7, 8, and 9). Each simple query contains 3 or 4 classes and each complex query has 6 or 7 classes. Since this set of queries has a large variety of query patterns, we feel that it can represent a general application of this database. The number in parentheses beside each class number is the measured processing cost of the class when the query is actually executed. We can calculate the processing cost of each class. For example, class 2 (Transcript) has been referenced twice in query 0 and query 8. The overall processing cost of class 2 in the set is the sum of the processing costs of class 2 in query 0 and query 8. Therefore, the overall processing cost of class 2 is 46.13. The calculated processing costs for all the classes referenced by the query set are shown in Table 1**.**

The overall processing cost of a class represents the minimal work that needs to be done for the set of queries if the class is assigned to a single processor. If we assign multiple classes to a processor, the load of the processor is the sum of the processing costs of the classes that are assigned to it. In order to achieve good performance, we want to distribute the load among the processors as evenly as possible. Load balancing is our main consideration for partitioning the database.

However, when we group two classes and put them on the same processor, the time for exchanging messages between these two classes can be drastically reduced. Therefore, we also want to group classes in a way so that the overall communication time can be reduced. The length of the longest path in a query is called the *query diameter***.**

If we reduce the diameter of a query, the overall communication time of the query will also be reduced. The query diameter can be reduced by grouping adjacent classes in the longest path and assigning them to the same processor. This is our secondary consideration for partitioning the database.

We combine the above two heuristics into the following method for partitioning a database. Since we want to evenly distribute the processing cost among the

processors, the number of groups of classes that we formed should be equal to the number of processors, assuming the number of classes is larger than the number of processors.

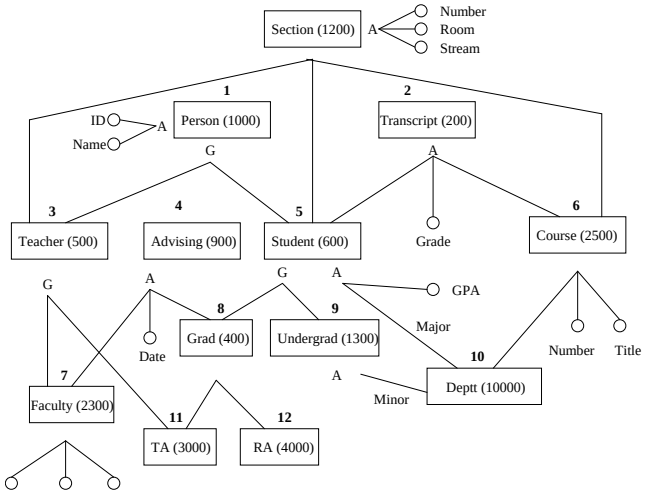


Fig1: The University database

The overall processing cost of each group should be close to the average processing cost among the processors. The average processing cost is called threshold cost. If some of the classes have processing cost that are larger than the average processing cost of the processors, we assign each of them to an empty group and will not assign any other class to these groups. The remaining classes should be distributed among the remaining processors as evenly as possible. Since the processing costs of the classes assigned to the single-class groups are above the threshold cost, the average processing costs of the remaining classes would be lower than the threshold cost. For this reason, we calculate a new threshold cost based on the costs of the remaining classes.

This new threshold cost is used as an upper limit for grouping classes in the first phase. When we group classes together, the total processing cost of the resulting group should not be larger than the new threshold cost. We start from the query with the largest diameter in the set and try to reduce the diameter by grouping two adjacent classes in the longest path. The two adjacent classes with the smallest combined processing cost will be considered. If the combined cost does not exceed the threshold cost, we group them together and use the combined cost as the processing cost of these two classes in all the queries. This step reduces the length of the longest path by 1. Then, we try to reduce the next longest path in the set by 1.

If there are multiple paths with the same length, we find a candidate pair of class for each path and choose the pair with the lowest combined cost to group. This process will continue until we cannot reduce the length of any path by grouping classes or the number of groups is equal to the number of the processors. In this phase, while we group classes to reduce the query diameters, the threshold cost is used to control the load in each group so that we can balance the load during the second phase of our heuristic method.

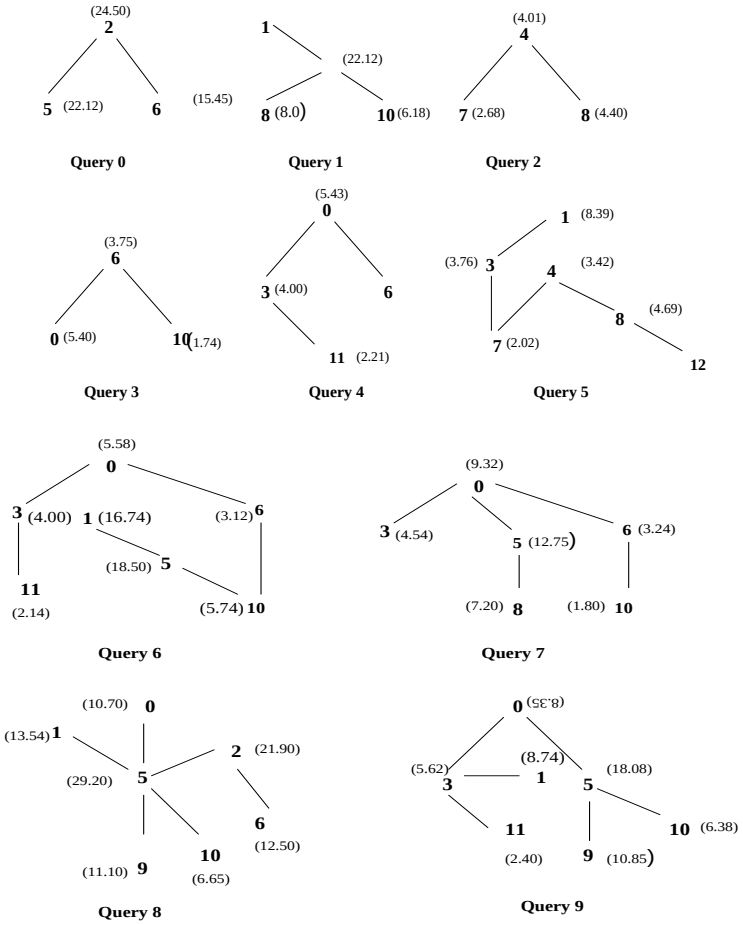


Fig 2: Sample Queries

Class No.	Cost
0	44.78
1	60.26
2	46.40
3	21.92
4	7.43
5	118.68
6	42.34
7	4.70
8	24.29
9	21.95
10	28.49
11	6.75
12	3.59

Table 1: Processing cost of each class

After we finish grouping classes for reducing query diameters, we need to reduce the number of groups to the number of processors in the system. In other words, we want to form the same number of clusters of groups as the number of processors. First, the groups are sorted based on

their processing costs. Then we assign the group with the largest cost to the first available cluster with the lowest cost and add the group's cost to the cluster's cost. By continuing this simple process, we can assign all groups to a fixed number of clusters having relatively close final costs among the clusters. Then, we can assign each cluster to a processor because we assume all the processors are the same and they are fully connected.

4. An Example

If we want to partition the university database for a 7-node system, we need to find a suitable set of queries to represent the application and measure the processing cost of each class in each query. An example is shown in Figure 2 Then, we calculate the overall processing cost of each class and the results are shown in Table 1. The next step is to find the threshold cost. Since the total processing cost of all classes is 434.59 and the average processing cost of the 7 processors is 62.08, the cost of class 5 is too large for it to be considered in the following procedure. Therefore, we just assign class 5 to a processor and drop it from further consideration. We also re-calculate the average cost of the remaining 6 processors and it is 52.65. This is the new threshold cost.

Among the 10 queries, query 6 has the longest diameter of 6. The adjacent classes 3 and 11 have the smallest combined cost (29.06) in the longest path in the query. We group them together. Now, queries 6 and 5 both have a diameter of 5. We check the longest path in query 6 and cannot find two adjacent classes that have a combined cost lower than the threshold cost. In query 5, we find classes 4 and 7 can be grouped together. By continuing this process, we find that the groups with their cost in parentheses are as follows: 5 (118.72); 1 (60.90); 9 and 10 (50.83); 2 (46.13); 0 (45.26); 6 (43.01); 4, 7, 8, and 12 (40.68); 3 and 11 (29.06).

We assign the 7 largest groups to the 7 empty clusters. Then, we assign the next group (in this case 3 and 11) to the lowest cluster. After we finish all the assignment, we have the following clusters: class 5; classes 3, 11, 4, 7, 8, and 12; class 1; classes 9 and 10; class 2; class 0; class 6. The heuristics presented here along with some other methods will be evaluated in the following chapter.

5. Evaluating Partition Heuristics

This method performs better than LB and OCPN methods of partitioning. The LB heuristic method does not try to reduce the query diameters in the query set. It directly goes to the second step and tries to balance the load. The one-class-per-node partitioning method (OCPN) assigns only one class to a node. Partition of our method performs better than the LB and OCPN methods when the communication delay increases. This means the heuristics used for partition the database is a good one.

6. Conclusion

We have proposed a heuristic method for partitioning the database. The database is partitioned for a specific application the processing requirement of which is represented by a set of queries. By analyzing the queries and the system characteristics, we can partition the database to suit the application. This heuristic method first uses a threshold cost as a guide to group small classes so that the query diameters can be reduced. Then, it tries to evenly distribute the cost among all the processors. This heuristic method is based on the overall processing cost of each class referenced in a query. We measure the overall CPU time and IO time used for processing a class to represent the processing cost of that class in the query. When we consider the set of queries, we take the sum of the processing costs for the same class in all queries to represent the total processing cost for that class.This method performs better than other partitioning methods e.g. LB and OCPN if the communication delay is long.

7. References

[1] J. R. Cheng, and A. R. Hurson, “Effective clustering of complex objects in object-oriented databases”, in ACM SIGMOD International Conference on Management of Data, Denver, CO, 1991, pp. 22-31.

[2] G. Copeland, and S. Khoshafian, “A decomposition storage model”, in ACM SIGMOD International Conference on Management of Data, Austin, TX, 1985, pp. 268-279.

[3] G. Copeland, W. Alexander, E. Boughter and T. Keller, ”Data placement in Bubba”, in ACM SIGMOD International Conference on Management of Data, Chicago, IL, 1988, pp. 99-108.

[4] D. J. DeWitt, R. Gerber, G. Graefe, M. Heytens, K. Kumar and M. Muralikrishna, ”GAMMA-A high performance dataflow database machine”, in 12th International Conference on Very Large Data Bases, Kyoto, Japan, 1986, pp. 228-237.

[5] D. J. DeWitt, S. Ghandeharizadeh, D. A. Schneider, A. Bricker, H. I. Hsiao and R. Rasmussen, “The Gamma database machine project”, IEEE Transactions on Knowledge Data Engg., Vol. 2, No. 1, 1990, pp. 44-62.

[6] L. W. Dowdy, and D. V. Foster, “ Comparative models of the file assignment problem” , ACM Computing Survey, Vol. 14, No. 2, 1992, pp. 287-313.

[7] S. Ghandeharizadeh, and D. J. DeWitt, “ Hybrid-range partitioning strategy: A new declustering strategy for multiprocessor database machines”, in 16th International Conference on Very Large Data Bases, Brisbane, Australia, 1990, pp. 481-492.

[8] G. Ghandeharizadeh, and D. J. DeWitt, “A multiuser performance analysis of alternative declustering

strategies", in 6th International Conference on Data Eng., Los Angeles, CA, 1990, pp. 466-475.

[9] S. Khoshafian, G. Copeland, T. Jagodits, H. Boral and P. Valduriez. "A query processing strategy for the decomposed storage model", in 3rd International Conference on Data Engg., Los Angeles, CA, 1987, pp. 636-643.

[10] S. Khoshafian, P. Valduriez and G. Copeland, " Parallel query processing for complex objects", in 4th International Conference on Data Engg., Los Angeles, CA, 1988, pp. 202-209.

[11] H. Lam, S. Y. W. Su, F. L. C. Seeger, C. Lee and W. R. Eisenstadt, "A special function unit for database operations within a data-control system", in International Conference on Parallel Processing, Chicago, IL, 1987, pp. 330-339.

[12] K. Shannon, and R. Snodgrass, "Implementing Persistent Object Bases: Principles and Practice", Morgan Kaufmann Publishers, Palo Alto, CA., 1991, pp. 389-402.

[13] M. M. Tsangaris, and J. F. Naughton, "A stochastic approach for clustering in object bases", in ACM SIGMOD International Conference on Management of Data, Denver, CO, 1991, pp. 12-21.

[14] P. Valduriez, "ACM Transactions on Database System", Vol. 12, No. 2, 1987, pp. 218-246.

Sudesh Rani got his Ph.D. degree in Computer Science from the Kurukshetra University, Kurukshetra, India, in 2009, on "Algebraic query processing and parallelism in databases". Her areas of interest are data mining, parallel databases, query processing in databases etc. Presently she is working as Asstt. Professor in Computer Science at Govt. College, Hisar, Kurukshetra University, Kurukshetra, India

A Review of Data Mining Classification Techniques Applied for Diagnosis and Prognosis of the Arbovirus-Dengue

A.Shameem Fathima[1] ,D.Manimegalai[2] and Nisar Hundewale [3]

[1] **Department of Computerscience and Engineering , Manonmanium Sundaranar University , Tirunelveli, Tamilnadu, India**

[2] **Department of Information Technology, National Engineering College, Kovilpatti,Tamilnadu,India**

[3] **Department of Computerscience and Information Technology, Taif University, Saudiarabia**

Abstract

Chikungunya (CHIK) virus, similar to Dengue pose a serious threat in Tropics, because of the year-round presence of Aedes mosquito vectors .The use of machine learning techniques and data mining algorithms have taken a great role in the diagnosis and prognosis of many health diseases. But a very few work has been initialized in this arboviral medical informatics. Our focus is to observe clinical and physical diagnosis of chikungunya viral fever patients and its comparison with dengue viral fever. Our project aims to integrate different sources of information and to discover patterns of diagnosis, for predicting the viral infected patients and their results. The scope is mainly in the classification problem of these often confused arboviral infections. This study paper summarizes various review and technical articles on arboviral diagnosis and prognosis. In this paper we present an overview of the current research being carried out using the data mining techniques to enhance the arboviral disease diagnosis and prognosis. This paper is not intended to provide a comprehensive overview of medical data mining but rather describes some areas which seem to be important from our point of view for applying machine learning in medical diagnosis for our real viral dataset.

Keywords: *Data Mining, Medical data, Machine learning algorithms, Diagnosis, Arbovirus.*

1. Introduction

Presently, in most parts of the Tropics, epidemics are near peak transmission before they are recognized and confirmed as viral infection. By then it is too late to generally implement effective preventive measures that could have an effective impact on transmission and thus on the course of the epidemic. Therefore the surveillance for Dengue/ Chikungunya should be proactive. This proactive surveillance system will permit prediction of Dengue / Chikungunya outbreak. The most important component of this system will also permit to differentiate whether the illness is Dengue or Chikungunya, as the initial symptoms are similar in both the disease. The objective of these predictions is to assign patients to either a “Dengue” group or a “Chikungunya” group or “any other infection” and to handle mystifying cases for the viral disease. Thus, arboviral diagnostic and prognostic problems are mainly in the scope of the widely discussed classification problems. These problems have paved a new face to many researchers in computational intelligence, data mining, and statistics fields.

Medical Informatics is generally clinical and/or biological in nature, and data driven statistical research has become a common complement. Predicting the outcome of a disease is one of the most interesting and challenging tasks where to develop data mining applications. As the use of computers powered with automated tools, large volumes of medical data are being collected and made available to the medical research group. As a result, Knowledge Discovery in Databases (KDD), which includes data mining techniques, has become a popular research tool for medical researchers to identify and exploit patterns and relationships among large number of variables, and made them able to predict the outcome of a disease using the historical cases stored within datasets. The objective of this study is to summarize various review and technical articles on diagnosis and prognosis of arboviral diseases. It gives an overview of the current research being carried out on various viral datasets using the data mining techniques to enhance the arboviral diagnosis and prognosis.2. Arboviral Infections –An Overview of Dengue and Chikungunya

2. Arboviral Infections-An Overiew of Dengue and Chikungunya

Chikungunya is a disease caused by the arbovirus that shares the same vector with dengue virus. Thus, in dengue-endemic region, Chikungunya is also a significant cause of viral fever causing outbreaks associated with severe morbidity. The symptoms of CHIKV infection are quite similar to those caused by many other infectious agents in the endemic areas. One particular difficulty in identifying CHIKV infection is its overlapping distribution with dengue viruses. It has been postulated that many cases of dengue virus infection are misdiagnosed and that the incidence of CHIKV infection is much higher than reported. Comprehensive study has not been undertaken to determine the clear picture of CHIKV infection and its comparison with dengue. Therefore, the present study was undertaken to diagnose Chikungunya infection in clinically suspected dengue patients and the vice–versa presented to The King Institute of Preventive Medicine, Chennai and in hospitals for diagnosis.

3. Knowledge Discovery and Data Mining

This section provides a preface to knowledge discovery and data mining. We provide the various analysis tasks that can be goals of a discovery process and lists methods and research areas that are challenging in solving these analysis tasks.

3.1. The Knowledge Discovery Process (KDD)

KDD is the process of extracting high-level knowledge from low-level data. Therefore, KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and KDD are often treated as the same words but in real data mining, it is an important step in the KDD process. The KDD process is often viewed as a multidisciplinary activity that encompasses techniques such as machine learning. The KDD process is interactive and iterative, involving numerous steps [1] such as

(1) Data cleaning: also called data cleansing, is a phase in which noise data and irrelevant data are removed from the collection.
(2) Data integration: is a stage in which heterogeneous multiple data sources are combined to a common source.
(3) Data selection: at this step, the data related to the analysis is decided on and retrieved from the data collection.
(4) Data transformation: also called as data consolidation, is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
(5) Data mining: it is the crucial step in which knowledgeable techniques are applied to extract potentially useful patterns.
(6) Pattern evaluation: this step, strictly interesting patterns representing knowledge are recognized based on the given measures.
(7) Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user.

3.2. Data Mining Process

Data mining is the process of selecting, exploring and modeling large amounts of data in order to discover unknown patterns or relationships which provide a clear and useful result to the data analyst [2]. There are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on available data. In the context of the data mining tasks, diagnosis and prognosis are to discover knowledge necessary to interpret the gathered information. In some cases this knowledge is expressed as probabilistic relationships between clinical features and the proposed diagnosis or prognosis. In other cases, the system is designed as a black-box decision maker that is totally unconcerned with the interpretation of its decisions. Finally, in yet other cases, a rule-based representation is chosen to provide the physician with an explanation of the decision. The latest is the most convenient way for physician to express their knowledge in medical diagnosis. Thus, the major challenge presented by medicine is to develop technology to provide trusted hypotheses based on measures which can be relied upon in medical research and clinical hypothesis formulation.

Data mining involves some of the following key steps [3] -
(1) Problem definition: The first step is to identify goals.
(2) Data exploration: All data needs to be consolidated so that it can be treated consistently.
(3) Data preparation: The purpose of this step is to clean and transform the data for more robust analysis.
(4) Modeling: Based on the data and the desired outcomes, a data mining algorithm or combination of algorithms is selected for analysis. The specific algorithm is selected based on the particular objective to be achieved and the quality of the data to be analyzed.
(5) Evaluation and Deployment*:* Based on the results of the data mining algorithms, an analysis is conducted to determine key conclusions from the analysis and create a series of recommendations for consideration.

4. Data Mining Classification Methods

Classification is the most frequently used data mining task with a majority of the implementation of Bayesian classifiers, neural networks, and SVMs (Support Vector Machines). A myriad of quantitative performance measures were proposed with a predominance of accuracy, sensitivity, specificity, and ROC curves. The latter are usually associated with qualitative evaluation.

Classification maps the data in to predefined targets. It is a supervised learning as targets are predefined. The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, the classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. The commonly used methods for data mining classification tasks can be classified into the following groups [4].

4.1. Decision Trees (DT's)

A decision tree is a tree where each non-terminal node represents a test or decision on the considered data item. Selection of a certain branch depends upon the outcome of the test. To classify a particular data item, we start at the root node and follow the assertions down until we reach a terminal node (or leaf). A decision is made when a terminal node is approached. Decision trees that use recursive data partitioning can also be interpreted as a special form of a rule set, characterized by their hierarchical organization of rules.

4.2. Support Vector Machine (SVM)

Support vector machines (SVM) are based on statistical learning theory and belong to the class of kernel based methods. SVM is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. Such a hyper plane is called the optimal hyper plane. A set of instances that is closest to the optimal hyper plane is called a support vector. Finding the optimal hyper plane provides a linear classifier. SVMs are well suited to dealing with interactions among features and redundant features.

4.3. Genetic Algorithms (GAs) / Evolutionary Programming (EP)

Genetic algorithms and evolutionary programming are algorithmic optimization strategies that are inspired by the principles observed in natural evolution. Genetic algorithms and evolutionary programming are used in data mining to formulate hypotheses about dependencies between variables, in the form of association rules or some other internal formalism.

4.4. Fuzzy Sets

Fuzzy sets form a key methodology for representing and processing uncertainty. Fuzzy sets constitute a powerful approach to deal not only with incomplete, noisy or imprecise data, but may also be helpful in developing uncertain models of the data that provide smarter and smoother performance than traditional systems.

4.5. Neural Networks

Artificial neural networks were recently the most popular artificial intelligence-based data modeling algorithm used in clinical medicine. Neural networks (NN) are those systems modeled based on the working of human brain. As the human brain consists of millions of neurons that are interconnected by synapses, a neural network is a set of connected input/output units in which each connection has a weight associated with it. The network learns in the learning phase by adjusting the weights so as to be able to predict the correct class label of the input. Neural networks may be able to model complex non-linear relationships, comprising an advantage over simpler modeling methods like the Naïve Bayesian classifier or logistic regression.

4.6. Rough Sets

The fundamental concept behind Rough Set Theory is similar to the Fuzzy set theory .The Difference is that the uncertain and imprecision in this approach is expressed by a boundary region of a set. Every subset defined through upper and lower approximation is known as Rough Set. Rough set is defined by topological operations called approximations, thus this definition also requires advanced mathematical concepts. They are usually combined with other methods such as rule induction, classification, or clustering methods.

5. Data Mining Classification Methods in use for the Data Mining of Arbovirus-Dengue

Clinical diagnosis of Dengue/ Chikungunya infection helps in predicting the viral cases. Suspected dengue case is defined as an acute febrile illness characterized by frontal headache, retro-ocular pain, muscle and joint pain, and rash (WHO, 2006). Besides the description of clinical symptoms, there are also clinical laboratory tests that are useful in the diagnosis of dengue. These clinical tests include a complete blood cell count (CBC), especially the

white blood cell count (WBC), platelet count and haematocrit levels. The results obtained from these methods are used to recognize the patterns which are aiming to help the doctors for classifying the malignant and benign cases. There are various data mining techniques, statistical methods and machine learning algorithms that are applied for this purpose. This section consists of the review of various technical and review articles on data mining techniques applied in arboviral dengue diagnosis.

In [5] Hani M. Aburas, B. Gultekin and Murat Sari predicted the dengue confirmed cases by using Artificial Neural Networks (ANNs). The model created by the authors were from 14,209 dengue reported confirmed-cases .They have taken many physical parameters such as mean temperature, mean relative humidity and total rainfall. Their prediction model has shown to be very effective processing systems for modeling and simulation in the dengue confirmed-cases data assessments as they did not use time information in building the model.

In [6] Janaína Gomide et al proposed a dengue surveillance approach that is a weekly overview of what is happening in each city compared with the weeks before. They construct a highly correlated linear regression model based on four dimensions: volume, location, time and content. Specifically, they showed that Twitter can be used to predict, spatially and temporally, dengue epidemics by means of clustering.

In [7], Silvia Rissino and Germano Lambert-Torres have used a Rough Set approach for the elimination of redundant data and the development of a set of rules that it can aid the doctor in the elaboration of the dengue diagnosis. From the dataset they roughly used for analysis they observed that patients with characteristics of all same attributes cannot be classified neither with dengue nor without dengue, but with only the decision attribute (dengue) not being identical and generates an inconclusive diagnosis for dengue.

In [8],Benjamin M. Althouse, Yih Yng Ng and Derek A. T. Cummings provided a comparison by analyzing dengue data from Singapore and Bangkok. Among the three models to predict incidence, SVM models outperformed logistic regression in predicting periods of high incidence. They found that the AUC for the SVM models using the 75th percentile cutoff is 0.906 in Singapore and 0.960 in Bangkok.

In [9], Ana Lisa V. Gomes et al, presented and implemented the novel application of the support vector In [9], Ana Lisa V. Gomes et al, presented and implemented the novel application of the support vector machines (SVM) algorithm to analyze the expression pattern of 12 genes in peripheral blood mononuclear cells (PBMCs) of 28 dengue patients (13 DHF and 15 DF) during acute viral infection. They achieved the highest accuracy of ~85% with leave-one-out cross-validation. However, their approach had a drawback that experimental investigation was s necessary to validate their specific roles in dengue disease.

In [10], Fatimah Ibrahim, Mohd Nasir Taib, Wan Abu Bakar Wan Abas, Chan Chong Guan and Saadiah Sulaiman developed a prediction system based prediction solely on the clinical symptoms and signs. Their system uses the multilayer feed-forward neural networks (MFNN) and is able to predict the day of defervescence in dengue patients with 90% prediction accuracy.

In [11], Madhu.G, G.Suresh Reddy and Dr.C.Kiranmai presented an intelligent approach to dengue data analysis with rough sets for the elimination of redundant data and development of set of rules that can help medical practitioners in patient's diagnosis. They processed the data based on the lower and upper approximations and theory was defined as a pair of the two crisp sets to the approximations.

In [12], Sree Hari Rao and Suryanarayana U Murthy developed a novel efficient classification algorithm designated as VB Classif 1.0 which is utilized to classify nearly 10,000 records with 94% accuracy. This tool has performed better than the well known K-Nearest neighborhood (KNN) algorithm with different sizes of train data.

In [13], Fatimah Ibrahim, M. I Mohamad, S. N. Makhtar and J. Ibrahim, developed a rule based expert system to classify three types of risk-higher, lower and no risk group among the dengue infections using bioelectrical impedance analysis (BIA).The classification process was done according to gender, reactance value of the BIA and 'day of fever' on a daily basis diagnosis. Their system successfully classified the risk in dengue patients noninvasively with total classification accuracy of 66.7%.

In [14], F. Ibrahim, T. Faisal, M. I. Mohamad Salim and M. N. Taib, used bioelectrical impedance analysis (BIA) and artificial neural network (ANN) to analyze the data of nearly 223 healthy subjects and 207 hospitalized dengue patients. Four parameters were used for training and testing the ANN which are day of fever, reactance, gender, and risk group's quantification. Their Best ANN architecture trained with the steepest descent back propagation with momentum algorithm obtained the prediction risk

classification accuracy of 95.88% for high risk and 96.83% for low risk groups.

In [15], the authors aim to create a clinical Data ware house for quick retrieval of reliable information on the viral diseases and their preventive measures at times of need. They claim that their ongoing work will be a boon to the researchers, academicians, Doctors, Health workers and Govt. servants and all for handy planning.

In [16], the authors employed decision tree algorithm to classify dengue infection levels into 4 groups (DF, DHF I, DHF II, and DHF III) and achieved an average accuracy is 96.50 %. The authors have compared their performance in term of false negative values to WHO and some researchers and found that that their research outperforms those criteria.

6. Proposed Work

There are many strains of Arbovirus. Our Research focus is on a particular variety. The main goal of the research is to first and foremost analyze the data from the surveys and to judge whether it is suitable to be analyzed with the use of the data mining methods. The second step is to evaluate several data mining algorithms in terms of their applicability to this data. Finally, an attempt is to be made to deliver some tangible medical knowledge extracted by the methods.

The analyses performed within this research are based on the data from the King Institute of Preventive Medicine and surveys filled out by patients and cards filled out by doctors from different hospitals. Data is extracted by using a standardized data collection form and is analyzed using R project version 2.12.0.

In summarizing we can separate the following methodological steps [17]:
1. Collecting and getting acquainted with a number of classification algorithms (e.g. data mining environment).

2. Reviewing the data set (e.g. a part of a patient health records).

3. Separating appropriate algorithms suitable for the data set.

4. Testing the full data set on selected number of classification algorithms, containing their default parameter values.

5. Selecting the best algorithms to use for further experiments.

6. Training the selected algorithms on reduced data set, by removing the attributes that appeared to be uninformative in building and visualizing the decision trees.

7. Modifying algorithms' default parameter values. Using the optimal data set formed for each algorithm of the most useful data identified in step 6.

8. Evaluating the results.

9. Randomizing the data set.

10. Performing steps 6 and 7 on randomized data set.

11. Evaluating and comparing results as well as algorithms performance.

These are the steps we have planned to perform with our data mining environment and data sets as well.

The choice of the R project [18] as the computational platform stems from its popularity and thus critical mass, ease of programming, good performance, and an increasing use in several fields, such as bioinformatics and finances, among others. Correlations between mortality and symptoms, physical examination findings, or laboratory findings at admission are examined using logistic regression, where appropriate. Correlations between the development of respiratory failure and the aforementioned patient characteristics are examined similarly.

Several data mining models are built using decision trees, clusters, neural networks, logistic regressions, association rules and Naive Bayes. However, prior to this, the dataset has been pre-analyzed. This was to see how the attributes are represented in terms of their values to determine the initial input set of attributes. This was followed by the analyses.

The experiment within this research is to be conducted according to a defined formula. For each of the chosen data mining algorithms a set of models is built. Each of them is generated for a different parameters setting. These parameters vary from one method to another thus each of the models is treated individually. This means that one algorithm can have more models built than others.

The dataset has been split into two subsets: training and testing. The training dataset contains both discrete and continuous attributes. Some of the algorithms in the R project require the input attributes to be discrete only, though. Thus the models have been divided into two groups with respect to this facet. The first group contains those that accept continuous attributes: decision trees,

clustering, neural networks and logistic regression. The other comprises algorithms which require the attributes to be discrete: association rules and Naive Bayes. After the models have been constructed, they undergo the evaluation step. Their performance is measured with the use of lift charts and classification matrices. Afterwards, from each of the methods the best model (the best parameters setting) is chosen for the ultimate comparison of the algorithms. Finally, the best model emerges. [19][20]

This interdisciplinary research requires new tools and methods for searching, retrieving, manipulating and integrating data from multiple sources to generate, validate and apply new public health models. The resulting challenges are magnified in that the data and processes that we study are dynamic. When the system is developed, the flexibility of the design will make data invaluable [21] for retrospective analyses of public health problems that have not been amenable to previous analyses.

7. Conclusion

The presented discussion on knowledge extraction from medical databases is merely a short summary of the ongoing efforts in this area. It does, however, point to interesting directions of our research, where the aim is to apply hybrid classification schemes and create data mining tools well suited to the crucial demands of medical diagnostic systems. It is proposed to develop a substantial set of techniques for computational treatment of these data. The approaches in review [22] are diverse in data mining methods and user interfaces and also demonstrate that the field and its tools are ready to be fully exploited in biomedical research.

Acknowledgments

A heartfelt gratitude is expressed by the authors of this paper to the Department of Virology, king Institute of Preventive Medicine and **R**esearch, Chennai, India, for releasing the viral data for research and education.

References

[1] Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R. 1996. Advances in Knowledge Discovery and Data Mining. Menlo Park, Calif.: AAAI Press.

[2] P. Giudici, Applied Data Mining Statistical Methods for Business and Industry, Wiley & Sons, 2003.

[3] The Data Mining Process.[Online]. http://publib.boulder.ibm.com/infocenter/db2luw/v9/index.jsp?topic=/com.ibm.im.easy.doc/c_dm_process.html .

[4] Han J. and Kamber M., Data Mining: Concepts and Techniques, 2nd ed., San Francisco, Morgan Kauffmann Publishers,2001.

[5] Hani M.Aburas,B.Gultekin Cetiner and Murat Sari, Dengue confirmed-cases prediction:A neural network model, Expert Systems with Applications: An International Journal , Volume 37 Issue 6, June, 2010

[6] Janaína Gomide, Adriano Veloso, Wagner Meira Jr., Virgílio Almeida, Fabrício Benevenuto, Fernanda Ferraz and Mauro Teixeira Dengue surveillance based on a computational model of spatio temporal locality of Twitter, Journal Web science 2011 ACM.

[7] Silvia Rissino and Germano Lambert-Torres, Rough Set Theory – Fundamental Concepts,Principals, Data Extraction, and Applications, Data Mining and Knowledge Discovery in Real Life Applications, pp. 438, February 2009

[8] Benjamin M. Althouse,, Yih Yng Ng and Derek A. T. Cummings , Prediction of Dengue Incidence Using Search Query Surveillance, PLoS Negl Trop Dis. 2011 August; 5(8): e1258.

[9] Ana Lisa V. Gomes, Lawrence J. K. Wee, Asif M. Khan, Laura H. V. G. Gil, Eresto T. A. Marques, Jr, Carlos E. Calzavara-Silva and Tin Wee Tan ,Classification of Dengue Fever Patients based On Gene Expression Data Using Support Vector Machines, PLoS One. 2010; 5(6): e11267

[10]Fatimah Ibrahim, Mohd Nasir Taib, Wan Abu Bakar Wan Abas,Chan Chong Guan, Saadiah Sulaiman, A novel dengue fever (DF) and dengue haemorrhagic fever (DHF) analysis using artificial neural network (ANN), Computer Methods and Programs in Biomedicine, Volume 79 Issue 3, September, 2005,p273-281

[11]Madhu.G, G.Suresh Reddy and Dr.C.Kiranmai, Hypothetical Description for Intelligent Data Mining, International Journal on Computer Science and Engineering, Vol. 02, No. 07, 2010, 2349-2352

[12]Sree hari Rao Vadrevu,Suryanaryana U Murthy,A Novel Tool For Classification of Epidemiological Data of Vector Borne Diseases,Journal Of Global Infectious Diseases,Jan-Apr 2010.

[13]Fatimah Ibrahim, M. I Mohamad, S. N. Makhtar and J. Ibrahim Classification of Risk in Dengue Fever and Dengue Haemorrhagic Fever using Rule Based Expert System, IFMBE Proceedings, 2007, Volume 15, Part 3, 50-53.

[14] F. Ibrahim, T. Faisal, M. I. Mohamad Salim and M. N. Taib , Non-invasive diagnosis of risk in dengue patients using bioelectrical impedance analysis and artificial neural network, Medical and biological engineering and computing Volume 48, Number 11, 1141-1148.

[15]Dr.M. Usha Rani , M.Kalpana Devi , D.M. Mamatha , R.Seshadri,, Yaswanth Kumar.Avulapti, Clinical Data Warehouse on Insect Vector Diseases to Human of Andhra Pradesh ,International Journal of Computer Science and Information Security,2010,Vol 8,Issue 5,P 240-244.

[16]Daranee Thitiprayoonwongse, Prapat Suriyaphol, and Nuanwan Soonthornphisaj., Data Mining on Dengue Virus Disease, 13th International Conference on Enterprise Information Systems (ICEIS 2011), pp. 32-41, June 8-11, 2011

[17]Mertik M., Kokol P., Zalar B.Gaining Features in Medicine Using Various DataMining Techniques //Computational Cybernetics ICC C 2005,IEEE rdInternational Conference. – 2005. P. 21–24
[18] www.r-project.org.
[19]Rodríguez, A., Carazo, J.M. and A., Trelles O.; "Mining Association Rules from Biological Databases"; Journal of the American Society for Information Science and Technology; 56(493-504); 2005
[20]S. Sfakianakis, M. Blazantonakis, I. Dimou, M. Zervakis, M. Tsiknakis, G. Potamias, D.Kafetzopoulos, D. Lowe, Decision Support Based on Genomics: Integration of Data and Knowledge Driven Reasoning, International Journal of Biomedical Engineering and Technology, Special Issue on Decision Support in Medicine.
[21]Quinlan, J. R. (1983). Learning efficient classification procedures. In Machine Learning:An Artificial Intelligence Approach, ed. Michalski, Carbonnel & Mitchell. Tioga Press.
[22] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann.

A.Shameem Fathima is currently a Ph.D student in the Department of Computer Science and Engineering at Manonmanium Sundaranar University,India.She obtained M.E in Computer Science and Engineering from Crescent Engineering College affiliated to Anna University in 2004.She has 5 years of teaching experience in different academic institutions in India and abroad. She has a proven career record and has published many papers in conferences. Her focus of research is Data Mining.

Dr.D.Manimegalai is Professor and Head of the Department of Information Technology in National Engineering College**.** She had her BE & ME from Government College of Technology, Coimbatore and PhD from Manonmaniam Sundaranar University, Tirunelveli.Her Current areas of research interests include Medical Image Processing and Data Mining and Image Retrieval. She is a life member of Computer Society of India, Institution of Engineers, System Society of India and Indian Society for Technical Education

Dr. Nisar Hundewale received his Ph.D. in Computer Science from Georgia State University, USA. He has worked at National Institutes of Health (NIH), USA, as post-doctoral Fellow. Currently, he is an Assistant Professor and Associate Dean for Research at Taif University. His research interests are Algorithms, Machine Learning, Bioinformatics, Distributed Computing, and Networking. He is a great inspiration and shore up to young researchers.

Minimal Feature set for Unsupervised Classification of Knee MR Images

Ms. Rajneet Kaur[1] and Dr. Naveen Aggarwal[2]

[1]Assistant Professor, Sri Guru Granth Sahib World University, Fatehgarh Sahib

[2]Assistant Professor, University Institute of Engineering & Technology, Panjab University

Abstract

Knee scans is very useful and effective technique to detect the knee joint defects. Unsupervised Classification is useful in the absence of domain expert. Real Knee Magnetic Resonance Images have been collected from the MRI centres. Segmentation is implemented using Active Contour without Edges. DICOM, Haralick and some Statistical features have been extracted out. A database file of 704 images with 46 features per images has been prepared. Unsupervised Classification is implemented with clustering using EM model and then classification using different classifiers. Learning rate of 5 classifiers (ID3, J48, FID3new, Naive Bayes, and Kstar) has been calculated. At the obtained learning rate minimal feature set has been obtained for unsupervised classification of Knee MR Images.

Keywords: *Unsupervised Classification, Segmentation, Feature Extraction, Knee MR Images.*

1. INTRODUCTION

MRI is one of the latest medical imaging technologies. An MRI (Magnetic Resonance Images) scan is a radiology technique that uses magnetism, radio, electric waves and a computer technology to produce images of body structure. The Magnetic resonance imaging used for Knee scans is very useful and effective technique to detect the knee joint defects. It is a non-invasive method to take picture of knee joint and the surrounding images. Images are produced and analysis is done on one image at a time. Radiologist diagnose whether the image is normal or abnormal. He does not give collective analysis of many images at a time. In other words current medical technologies are not used for analysis purpose of multiple images and to give informative result about those images together at a time which can be helpful in future. This is what data mining used for. Along with this because of busy schedules of radiologists, it is difficult to go through large number of images every day. So an Unsupervised Classification of MRI image is used for discovering meaningful patterns and relationships that lie hidden within very large database in the absence of radiologist.

2. SEGMENTATION

Segmentation is the first step in the process of classification of images. Segmentation algorithms varies from edge based, region based and other thresholding techniques. In computer vision, segmentation refers to the process of partitioning a digital image into multiple segments (sets of pixels, also known as super pixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze [1]. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics.The result of image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image (see edge detection). Each of the pixels in a region is similar with respect to some characteristic or computed property, such as color, intensity, or texture. Adjacent regions are significantly different with respect to the same characteristic.

The segmentation techniques used is ‘Active Contour without edges’ by chen and vese. The active contour model by chen and vese [2] is used to detect the objects in the images using the technique called as curve evolution which was originally proposed by mumford-shah [3] function for segmentation and for defining level sets.

2.1 IMPLEMENTATION OF SEGMENTATION

Distance map of initial mask: Chan and Vese have used phi0 which is a distance map of initial mask in his work, but we can modify it to phi1, because converting image to double and then adding to phi calculation does not make any difference. So we can emit this parameter and shorten the equation.

phi0 = bwdist(m)-bwdist(1-m)+im2double(m)-.5;

phi1 = bwdist(m)-bwdist(1-m)-.5;

subplot(2,2,1); plot(phi0),title('phi0'),subplot(2,2,2); plot(phi1),title('phi1');

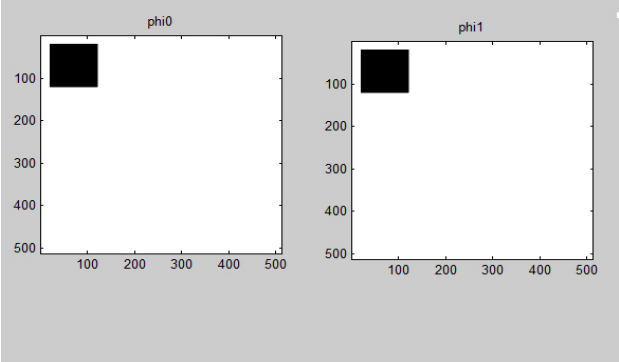


Figure 1: Distance Mask Of initial Mask

MRI images are larger in size. In order to process them it is needed to change the value of parameter μ if μ is small, then only smaller objects will be detected if μ is larger, then it will work for larger objects also or objects formed by grouping.

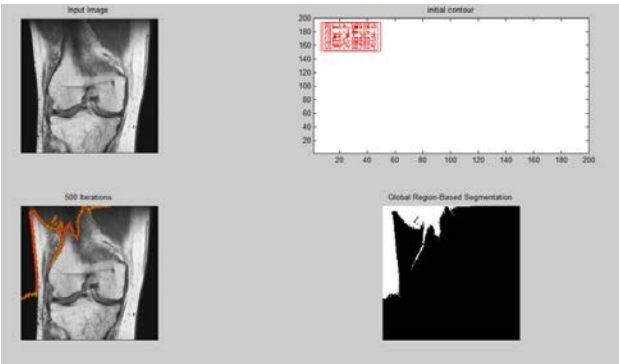


Figure 2: 'chan' or 'vector' method at dt=0.5

Increment the value of dt accordingly.

Dt = 2.5

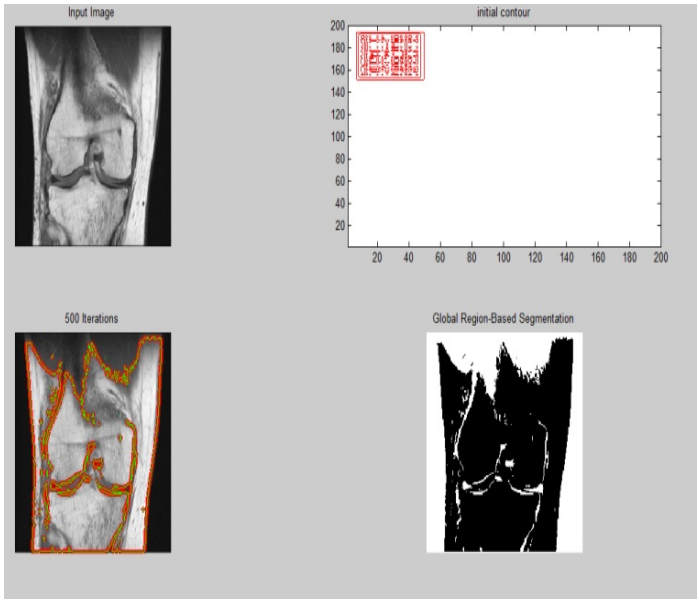


Figure 3: 'chan' or 'vector' method at dt=2.5

The output obtained by both 'chan' and 'vector' method is approximately same. But the difference in both the methods is this that if image is noisy then 'vector' method gives better result than the 'chan' method. So in the case of noisy images it is preferable to use 'vector' method.

2. 'Multiphase' method: In the case of 'multiphase method:

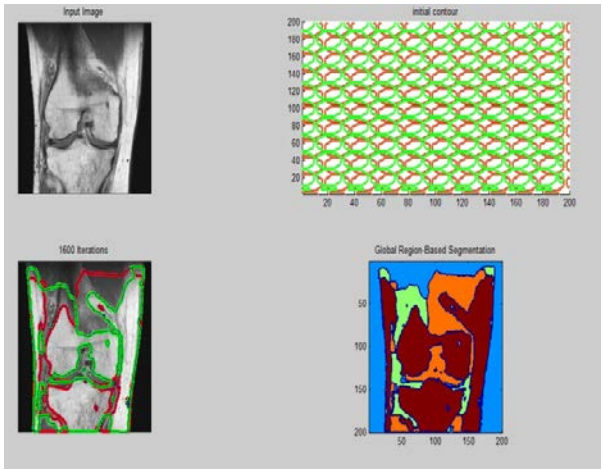
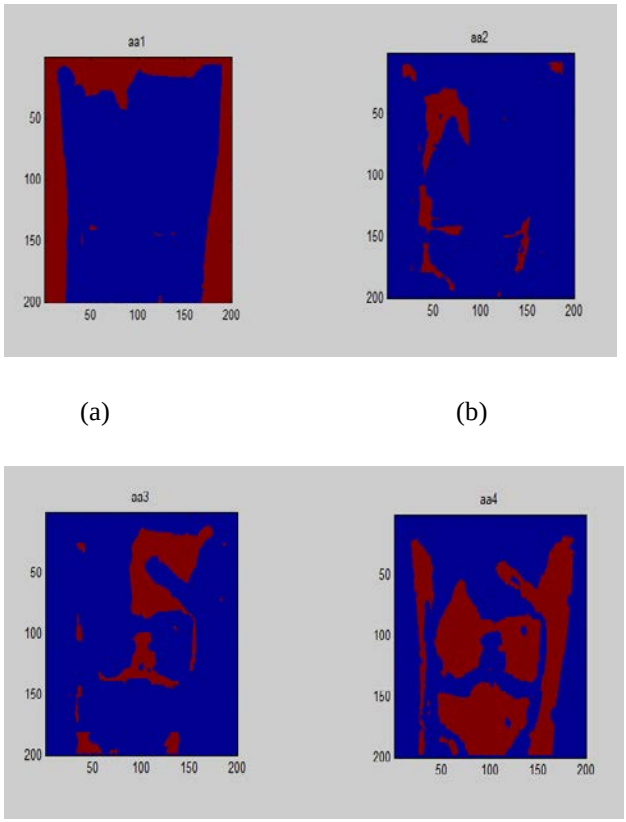


Figure 4: 'multiphase' method

Four regions aa1, aa2, aa3 and aa4 together form the complete segmented image. We can extract each region separately. By separating each region part we can get our region of interest without applying

any other technique on segmented image.



(a) (b)

(c) (d)

Figure 5: Sub regions of Active Contour without edge

3. FEATURE EXTRACTION

Large number of algorithms has been proposed for the extraction of features from knee MRI images. Texture analysis serve as a base for various feature description. Statistical, structural, spectral, filtering, histograms, transformation and many more methods are used for texture feature extraction. The global features capture the gross essence of the shapes while the local features describe the interior details of the trademarks. In the feature extraction part total 45 features have been extracted from Knee MRI images. Out of which 19 are DICOM images header features [4], 13 are haralick texture features [5] and rest are images statistical features [6]. DICOM header features are extracted out in order to check either all images have been taken under similar environmental conditions or not.

Features extracted:

Following 46 features per file comprises database file-

1.File Size	17.Flip Angle
2.Width	18.Rows
3.Height	19.Colums
4.Bit Depth	20.Angular Moment
5.PatientName	21.Contrast
6.Patient Birth Date	22.Correlation
7.Patient Sex	23.Entropy
8.Patient's Age	24.Inverse Difference Moment
9.Patient's Weight	25.Sum Average
10.Body part examined	26.Sum Variance
11.Slice Thickness	27.Sum Entropy
12.Image Frequency	28.Difference Average
13.Image Nucleus	29.Differnce Variance
14.Magnetic Field Strength	30.Differnce Entropy
15.Spacing between Slices	31.Infoiormation of Correlation1
16. Pixel Bandwidth	32.Infoiormation of Correlation2

4. UNSUPERVISED CLASSIFICATION

Classification is a process which is used to categorize the data (XML, images, text etc) into different groups ("classes") according the similarities between them [7]. Image classification is defined as the process to classify the pixels of images into different classes according to similarity.

In Unsupervised Classification, there is no expert present for prediction. To implement this firstly divide the data into cluster using any clustering approach and then apply classification algorithms which used the information of cluster not of any expert to classify the data [8]. For clustering we used 'EM' clustering algorithm. It is a method of finding the maximum likelihood of parameters in statistical model, where the model depends on unobserved latent variables. EM clustering is an iterative that alternates between performing an expectation (E) step, which computes the expectation of the log-likelihood evaluated using the current estimate for the

latent variables, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

Clustering has been implemented. The data is divided into different clusters and we saved the cluster assignment file in 'ARFF' format, whose last attributes shows the cluster assignment. The generated clustered file is used as input for classification in the next phase. Algorithms 'ID3', 'J48', 'FID3new', 'Naive Bayes' & 'Kstar' has been implemented and results are recorded and studied for analysis purpose. FID3new is an algorithm which is made by combining the two algorithms ID3 and FT, so it is a hybrid algorithm. In this the information gain and entropy measure of attributes is calculated according to ID3 algorithm and then put as input for the classification and the further classification is done according to FT algorithm. Improved results have been obtained on our dataset.

4.1 IMPLEMENTATION

Knee MRI scans has been collected and after image processing total 46 features have been extracted from the Knee MR images .A database file of 704 tuples and 46 attributes has been made in ASCII in CSV format, then conversion of this file to CSV file is done. CSV files are readable in Weka [9]. The generated CSV file is opened in Weka and then different processes like data cleaning, data processing and data transformation are applied on to the input database file. These steps act as pre-processing steps for the classification of data. Along with this attribute removal is also done. Some of the attributes like patient's name, patient weight, body part examined etc are removed as they contribute nothing in classification process. Classification is implemented using 'ID3', 'J48', 'FID3new', 'Naive Bayes' & 'Kstar' and results are recorded and studied for finding the minimal feature set for unsupervised classification.

In order to find out minimal features set for Knee MR Images in case of supervised classification, first step is to find out the learning rate of different algorithms. To find out the learning rate of different algorithms, the training is started from 1 % percentage split and keeps on increasing till 99% percentage split. Results of different algorithms have been recorded and analysed and interpretation has been done according to the analyses.

4.1.1 COMPARISON OF TP RATE VS PERCENTAGE SPLIT OF DIFFERENT ALGORITHMS FOR UNSUPERVISED CLASSIFICATION

Clustering is implemented on original database file with EM clustering algorithm. Output file is saved in ARFF format and given as input during classification. Training rate is started from 1% and gradually increased by 5 in each iteration till 99% of training. TP Rate of different algorithms has been calculated and plotted below. All algorithms behave differently according to their working.

It is observed that at 50% percentage split ID3 & FID3new gives TP Rate of 1 and gives constant value. All other algorithms also gives value near to 1 and get stabilized at this value. So we can say that 50% of training is required in case of unsupervised Classification. This is the minimum and required training which is must in order to get the proper and correct results. If we keep on increasing the value of training rate from 50 %, then there are almost same values obtained over all other percentage splits.

In case of FID3new after 50 % near about 90 % the TP rate goes below the value specified at 50% of percentage split. This is because of the over training. So we can say that 50 % is the required training in case of supervised classification.

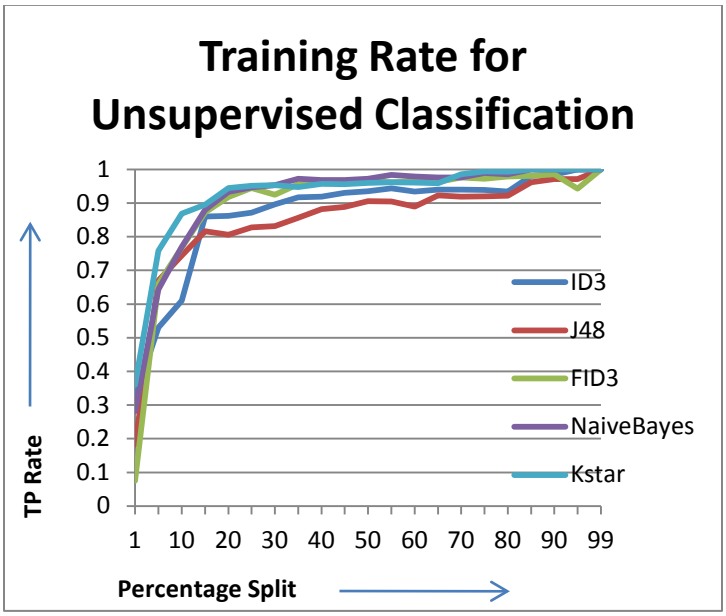


Figure 4.1: Training rate for unsupervised classification

4.1.2 COMPARISON OF FP RATE VS PERCENTAGE SPLIT FOR UNSUPERVISED CLASSIFICATION

FP Rate of different algorithms has been calculated and plotted below. In this graphs training rate is started from 1% and then we keep on increasing the training rate by 5 in each step and continue till 99 % of training. All algorithms behaves differently according to their working

It has been concluded from the above graph that at 50% percentage split ID3 & FID3new gives FP Rate of 0 and gives constant value. All other algorithms also gives value near to 0 and get stabilized at this value. So we can say that 50% of training is required in case of Supervised Classification. This is the minimum and required training which is must in order to get the proper and correct results. If we keep on increasing the value of training rate from 50 %, then there are almost same values obtained over all other percentage splits.

In case of FID3new after 50 % the FP rate sometime goes above the value specified at 50% of percentage split. This is because of the over training. So we can say that 50 % is the required training in case of supervised classification.

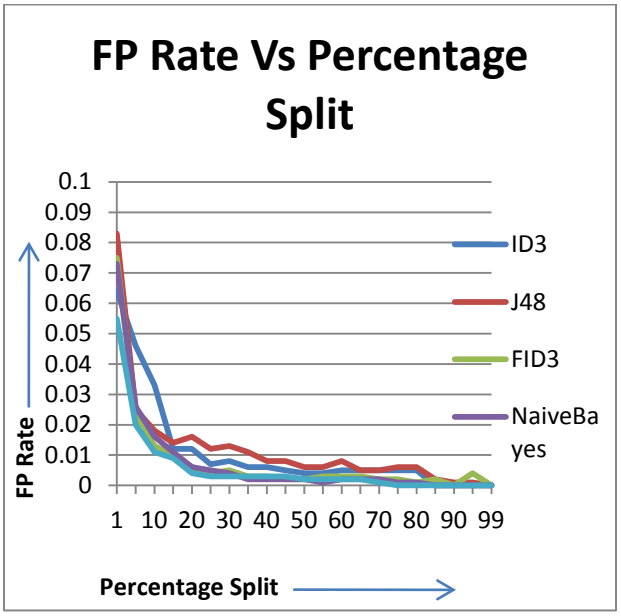


Figure 4.2: FP Rate Vs Percentage Split

In brief 50% of training is required in case KNEE Magnetic Resonance Images. At this percentage split of training minimal feature set for KNEE images has been obtained for unsupervised classification.

4.1.3 MINIMAL FEATURE SET FOR UNSUPERVISED CLASSIFICATION

Training of 50% is required in case of unsupervised classification. To find out minimal feature set, start the evolution from 2 attributes and increase the number by 2 in every step till all 41 attributes has not covered.

It has been observed from the plotted values that at 20 attributes all algorithms give maximum value of TP Rate. After 10 attributes ID3, FID3new and J48 gives constant value as they get stabilized, however the value of Naive Bayes and Kstar decreases. This is because of the over fitting of data. Over fitting occurs when the information available for proper classification is more than the required one. Other reason for over fitting is that during the training phase data is trained on different data, whereas its evaluation is done on some unknown data.

Most of the classifiers start memorizing the training data rather than to generalize them which results in over fitting of the data.

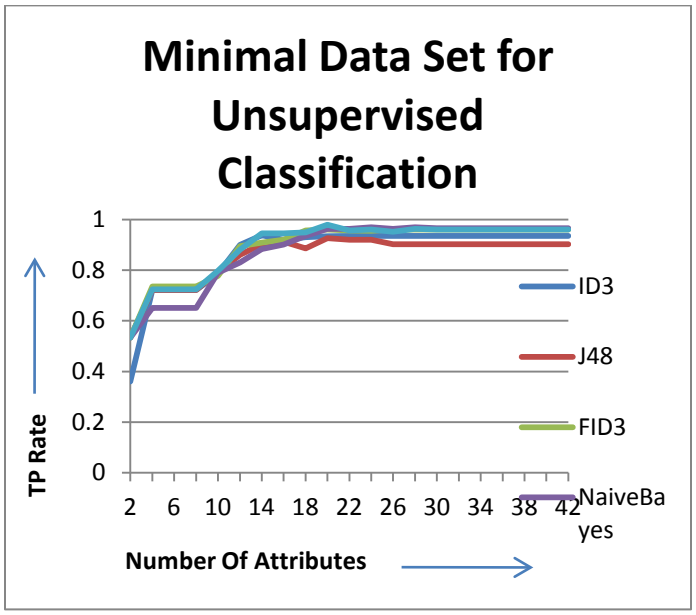


Figure 4.3: Minimal feature set for unsupervised classification

5. CONCLUSION

Real Knee MRI data have been collected from MRI canters. Segmentation is implemented using Active Contour without edges. It is easy to separate them out easily and can easily access the part containing cartilage thickness. In the next phase, total 46 features have been calculated and in the pre-processing 5 features which give the detail of patient's personal data have been removed. A database file consisting of 704 images with 41 lists of attributes is prepared and it used for classification process in next phase. Classification is implemented and performance of different parameters are compared using five algorithms 'ID3', 'J48', 'FID3new', 'Naive Bayes' & 'Kstar'. 'FID3new' is a hybrid algorithm, which is proposed in this work. . In unsupervised classification learning rate of different algorithms is calculated by starting the training from 1 % till 99 % and it has been concluded that minimum 50% of training is required in the case of unsupervised classification also. At this training rate minimal feature set has been calculated by taking minimum 2 features in starting and then increase the number two in each iteration till 42 features (one more feature that defines the cluster assignment). In case of unsupervised classification minimal feature set consist of 20 features and 'slice thickness' is the feature with highest priority. Classification is done using different algorithms. It has been concluded that 'FID3new' correctly classifies all instances and gives TP rate of 1 and Root Means Square's Error value 0. It classify the database on the base of feature 'Slice thickness' and divides them into four classes A, B, C & D. Where A= 0.9 mm, B=3, C= 4 and D= 6. The images coming under B & D class is classified as 'Normal' images and the images coming under A & D class is classified as 'Abnormal' images

6. FUTURE WORK

Only on 704 knee MR Images. Database can be extended and same methodology can be applied to the database containing images in thousands and many more. Only MRI Knee data has been used, the same approach can be extended to different medical imaging technologies like CT scan etc. Different segmentation algorithms can be used for segmentation. More features like Zernike moments etc can be calculated and feature set can be extended. Similarly for classification different combination of algorithms can be tried and results can be compared, if any improvements will be there, then it can be suggested.

7. REFERENCES

[1]Qi Luo,Wuhan,"Advancing Knowledge Discovery and Data Mining", Procedings of Knowledge Discovery and Data Mining,IEEE WKDD, pages 3-5, ISBN: 978-0-7695-3090-1, 2008

[2] T. Chan and L. Vese, "Active contours without edges," in IEEE Trans on Image Processing Vol 10,pages 266-278, ISBN: 1057-7149, 2001.

[3]D. Mumford and J. Shah, "Optimal approximation by piecewise smooth functions and associated variational problems", Commun. Pure Appl. Math, vol. 42, pages 577–685, 1989.

[4] Rosset, A. Spadola, L. Rati., "OsiriX: An Open-Source Software for Navigating in Multidimensional DICOM Images", JOURNAL OF DIGITAL IMAGING, vol 17, part 3, pages 205-216, ISBN: 0897-1889 , W B SAUNDERS CO, 2004.

[5] R. M. Haralick and K. Shanmugam, "Computer Classification of Reservoir Sandstones," IEEE Transactions on Geoscience Electronics, vol. 11, pages. 171-177, 1973.

[6] Brandt, S. Laaksonen, J. Oja, E., "Statistical Shape Features in Content-based Image Retrieval", International Conference On Pattern Recognition, vol 15; vol 2, pages 1062-1065, ISBN: 1051-4651,2000.

[7] Wu, X. Kumar, V. Ross Quinlan, J. Ghosh, J. Yang, Q. Motoda, H. McLachlan, G. J. Ng, A. Liu, B. Yu, P. S., "Top 10 algorithms in data mining", Knowledge and Information Systems, vol 14; number 1, pages 1-37, ISBN:0219-1377, Springer, 2008.

[8] Fayyad, U. M., G. P. Shapiro, P. Smyth.,"From Data Mining to Knowledge Discovery in Databases", AI Magazine, vol 17, number 3, pages 37-54, ISBN: 0738-4602, American Association of Artificial,1996.

[9] Holmes, G.; Donkin, A.; Witten, I.H., "WEKA: a machine learning workbench", Intelligent Information system, proceedings of second Australian and new Zealand conference, pages 357-361, ISBN: 0-7803-2404-8 , dec 1994

An Analysis of MIPS Group Based Job Scheduling Algorithm with other Algorithms in Grid Computing

S. Gomathi,
A.P,FXEC,
Tirunelveli,
Tamilnadu,India
Mobile: 9944866629

Dr.D.Manimegalai,
Prof & Head, IT Dept,NEC,
Kovilpatti,
Tamilnadu,India
Mobile: 9442636698

Abstract

Two major problems in grid computing applications are, resource management and job scheduling. These problems do occur due to distributed and heterogeneous nature of the resources. This paper introduces a model in job scheduling in grid computing environments. A dynamic scheduling algorithm is proposed to maximize the resource utilization and minimize processing time of the jobs. The proposed algorithm is based on job grouping. The results show that the proposed scheduling algorithm efficiently utilizes resources at its best and reduces the processing time of jobs.

Keywords- Grid computing; Job grouping; Job scheduling; Dynamic scheduling; First come first served (FCFS) algorithm.

1. INTRODUCTION

Grid computing refers to the cooperation of multiple processors and its aim is to use the computational power in the areas which need high capacity of the CPU. The Grid is concerned with the exchange of computer power, data storage, and access to large databases, without users searching for these resources manually. Grid computing is based on large scale resources sharing in an Internet. Computational Grids are emerging as a new computing paradigm for solving challenging applications in science, engineering, economics and econometrics [1]. Computational Grid can be defined as large-scale high-performance distributed computing environments that provide access to high-end computational resources. And also it is defined as a type of parallel and distributed system that enables the sharing, selection, and aggregation of geographically distributed autonomous resources dynamically at runtime depending on their availability, performance, capability, cost, and user's quality-of-service requirements.

Grid scheduling is the process of scheduling jobs over grid resources. A grid scheduler is in-charge of resource discovery, grid scheduling (resource allocation and job scheduling) and job execution management over multiple administrative domains. In heterogeneous grid environment with its multitude of resources, a proper scheduling and efficient load balancing across the grid can lead to improved overall system performance and a lower turn-around time for individual jobs. There are two types of scheduling namely static scheduling and dynamic scheduling in grid computing system. For static scheduling, jobs are assigned to suitable resources before their execution begin. For the dynamic scheduling, reevaluation is assigned to already taken assignment decisions during job execution.

In grid computing system, resources are not under the central control and can enter and leave the grid environment at any time. An effective grid resource management with good job and resource scheduling algorithm is needed to manage the grid computing system. In grid computing environment, there exists more than one resource to process jobs. One of the main challenges is to find the best or optimal resources to process a particular job in term of minimizing the job computational time. Optimal resources refer to resources having high CPU speeds and large memory spaces. Computational time is a

measure of how long that resource takes to complete the job.

In a Grid computing environment, the scheduler is responsible for selecting the best suitable machines or computing resources for processing jobs to achieve high system
throughput [2]. The scheduler must use coarse-grained jobs instead of light weight jobs so as to reduce communication and processing time. This paper focuses on grouping based job scheduling and how they are grouped as coarse grained jobs. The grouped jobs are allocated to resources in dynamic grid environment taking into account memory constraint, processing capabilities ,and the bandwidth of the resources.

This paper is organized as follows. In Section II, related work is surveyed, in section III basic grouping based job scheduling model is discussed, in section IV analyses experimental evaluation using GridSim toolkit [6] and section V concludes the paper with future work.

2. RELATED WORK

In the field of grid resource management and job scheduling, researchers have done much valuable work. Various algorithms have been proposed in recent years and each one has particular features and capabilities. In this section we review several scheduling algorithms which have been proposed in grid environment. Jobs submitted to a grid computing system need to be processed by the available resources. Best resources in terms of processing speed, memory and availability status are more likely to be selected for the submitted jobs during the scheduling process. Best resources are categorized as optimal resources.

A scheduling optimization method should consider the following two aspects, one is the application characteristics, and the other is the resource characteristics [6]. Taking the characteristics of lightweight job, into account there are some researches on the fine-grained job scheduling problem.

A dynamic job grouping-based scheduling algorithm groups the jobs according to MIPS of the available resources. This model reduces the processing and communication time of the job, but this algorithm doesn't take the dynamic resource characteristics into account and the grouping strategy may not utilize resource sufficiently [3].

A Bandwidth-Aware Job Grouping-Based scheduling strategy schedules the jobs according to the MIPS and bandwidth of the selected resource, and sends job group to the resource whose network bandwidth has highest communication or transmission rate. But, the strategy does not ensure that the resource having a sufficient bandwidth will be able to send the job group within required time [5].

Scheduling framework for Bandwidth-aware strategy schedules jobs in grid systems by taking of their computational capabilities and the communication capabilities of the resource's into consideration. It uses network bandwidth of resources to determine the priority of each resource. The job grouping approach is used in the framework where the scheduler retrieves information of the resources processing capability. The scheduler selects the first resource and groups independent fine-grained jobs together based on chosen resources processing capability. These jobs are grouped in such a way that maximizes the utilization of the resource's and reduces the total processing time. After grouping, all the jobs are sent to the corresponding resource's whose connection can be finished earlier which implies that the smallest request is issued through the fastest connection giving best transmission rate or bandwidth. However, this strategy does not take dynamic characteristics of the resources into account, and preprocessing time of job grouping and resource selection are also high [4].

The above analysis of various grouping based job scheduling strategy presents some of their advantages and disadvantages .However, there are some defects in the above scheduling algorithms. First, the algorithms doesn't take the dynamic resource characteristics into account. Second, the grouping strategy can't utilize resource sufficiently. And finally, it doesn't pay attention to the network bandwidth and memory size. To solve the problems mentioned above, an adaptive fine grained job scheduling mechanism is presented in this paper.

3. JOB SCHEDULING MECHANISM

The job scheduler is a service that resides in a user machine. Therefore, when the user creates a list of jobs in the user machine, these jobs are sent to the job scheduler for scheduling arrangement. The job scheduler obtains information about the available

resources from the Grid Information Service (GIS). Based on this information, the job scheduling algorithm is used to determine the job grouping and resource selection for grouped jobs. The size of a grouped job depends on the processing requirement length expressed in Million Instructions, Bandwidth expressed in MHz/s and Memory size requirement expressed in MB, expected execution time in seconds. As soon as the jobs are put into a group with a matching selected resource, the grouped job is dispatched to the selected resource for computation.

The grouping strategy should be based on the characteristics of resources. In grid computing, there are two approaches for obtaining dynamic resource characteristics for job execution. One is that a user directly searches the resources for job execution using an information service. The other is to use a resource manager. With a resource manager, users can obtain information about the grid through an interactive set of services, which consists of an information service that is responsible for providing information about the current availability and capability of resources. The resource monitoring mechanism used in the proposed algorithm belongs to the second one.

Grouping strategy is done based on the resource's status according to processing capabilities (in MIPS), bandwidth (in MHz/s), and memory size (in MB) of the available resources. After gathering the details of user jobs and the available resources, the system selects jobs in FCFS order to form different job groups. The scheduler selects resources in FCFS order after sorting them in descending order of their MIPS. Jobs are put into a job group one after another until sum of the resource requirements of the jobs in that group is less than or equal to the amount of resources available at the selected resource site. Here, only the processing capability and bandwidth are used to constrain the sizes of coarse-grained jobs, but we can easily join additional constraints .Then the fine-grained jobs can be grouped as several new jobs and these new jobs should satisfy the following formula:

1. MI (job_group_i) ≤MIPS (i)*tp(i)
2. FSG (job_group_i) ≤BW (i)*tc(i)
3. TMR (all_group) ≤TMA(all_resources)
4. tp>tc

In the above conditions, MI(job_group_i) is the processing capacity of the resource i which will be allocated to the jobgroupi , tp(i) is the expected job processing time, FSG(job_group_i) is the file_size (in Mb) of the jobgroupi at the resource i , tc(i) is the communication time, BW(i) is the bandwidth of resource i, TMR denotes the total amount of memory needed during the execution of the job j, TMA denotes the total amount of memory available.

Equation (1) specifies that the processing time of the coarse-grained job shouldn't exceed the expected time. The communication time of the grouped jobs should not exceed computation time of the grouped jobs and this is illustrated as (2) & (4). Equation (3) specifies that the memory size requirement of the jobgroup shouldn't exceed to the resource memory size. These are the constraints in job grouping.

This algorithm is divided into two parts. In the first part, the scheduler receives resource status using GIS. And, it sorts job list in descending order, and assigns a new ID for each job. In the second part after gathering the details of user jobs and the available resources, the system selects Jobs in FCFS order to form different job groups. The scheduler selects resources in FCFS order after sorting them in descending order of their MIPS. Jobs are put into a job group one after another until sum of the resource requirements of the jobs in that group is less than or equal to amount of resource available at the selected resource site.

In this way jobs are subsequently gathered or grouped one by one according to the resulting MIPS, Memory size and Bandwidth of the resource until the above conditions are satisfied. As soon as a job group is formed, the scheduler submits the grouped job to the corresponding resource for job computation setting the resource power to zero. After execution the job group, the results goes to the corresponding users and resource is again available to Grid system.

Algorithm:

Begin

Part 1: Initialization

Step 1.

Direct jobs to the Scheduler.

Step 2.

Direct Resource status to the scheduler

Step 3.

Sort joblist in descending order based on MIPS

Part 2: Job Scheduling

Step 1.

[Traverse Joblist] For i<-0 to joblistsize-1 do through step 2

Step 2.

[Traverse Grouplist] For j<-0 to joblistsize-1 do through step 3

Step 3

[Compare Processing Time] if MI(job_group_i)+jobi ≤ MIPS(i)*tp(i)

And

[Compare File Size] (jobgroup_ file_ sizej + job_ file_ sizei) / baud_ ratej <tp(i)) or MI(jobgroup_j) = 0,

And

[CompareMemory] jobi/MIPSj > job_file_size /baud_ratej)

Step 4 [Construct job group] add job i to job group j;

Step 5 [Loop Break] break;

Step 6 [End Compare] endif

Step 7 [Increment j] j++;

Step 8 [End Loop] endfor

Step 9 [Compare Status of i] if job i can't join any job_group then

Step 10 [Construct joblist2] add job i to joblist2;

Step 11 [Compare End] endif

Step 12 [Increment i] i++;

Step 13 [End Loop] endfor

Step14[Travese jobgrouplist_size] for i:<-0 to jobgrouplist_size-1

Step 15[Allocate jobgroup to resource] jobgroup I <- resource i;

Step 16[End Loop] endfor

Step 17[Compare size of joblist2] if joblist2_size<>0 then

Step 18[Assign joblist] joblist<-joblist2;

Step 19[Synchronize Process] wait a while;

Step 20[Get Resource status] get resource status from GIS;

Step 21[Receive jobgroup] receive computed jobgroup from resources;

Step 22 [Loop part2] repeat part2;

Step 23[Compare End] endif

Step24 [Receive Computed Jobgroup] receive computed job group from resources.

Step 25 [End of Algorithm] Ends

There are disadvantages in the above discussed algorithm. One of the major disadvantages is that there are some specific set of jobs that require only to a specific set of resources for assignment. For an instance, the job jx can be done only by the resource rx. Therefore the job jx cannot be assigned to any other resource. Another disadvantage is that as it is also possible that some job may require the processing capabilities of more than a resource. Either a parallel assignment or sequential assignment may be considered as a solution at times. In grid computing architecture dynamic scheduling, the resources are not at all under central control. A resource may enter and leave the environment at any time. The frequency of how frequently a resource enters to grid environment stays and moves away have to be carefully accounted. This statistics will help in placing the jobs in queue. The resource that most frequently enters into the grid can handle processing of jobs without much delay. A random based assignment could be a most ideal choice in the scheduling structure. Resource Manager gets information about the next entering resource into the grid environment. This in turn informs the job scheduler to regroup jobs that can be assigned to the entering resource. The job scheduler informs the list processor to regroup the jobs that can be assigned to the incoming resource. The group of jobs thus rescheduled will be made available to the resource. If there are n numbers of resources that a processor r can process, then the jobs within the group can be diverted to the resource in the FCFS pattern. The factors such as bandwidth, processing capabilities and memory size should be accounted in listing the jobs within a group (ordering). Jobs that require resource processing may be put on cycle till it finds a suitable resource entering into the grid.

4. EXPERIMENTAL EVALUATION

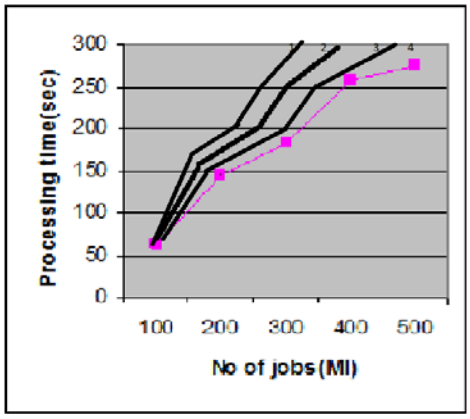
GridSim [6] has been used to create the simulation of grid computing environment. In this simulation, each resource is characterized by its MIPS, bandwidth and memory size. The jobs are characterized by their amount of computations, expected execution time, memory-size requirement and expected transfer time. In this experiment, jobs and resources are randomly generated and the number of jobs varies from 100 to 500. Jobs in

different groups are given different amount of execution time. The processing time is taken into account to analyze the feasibility of the proposed scheduling algorithm. Our algorithm can reduce the execution time and also the job completion success rate is high.

Table1: Job processing table

SNO	No.of Jobs	Processing Time(FCFS)	Processing Time(GBDJS)
1	100	55	55
2	200	200	160
3	300	280	220
4	400	380	260
5	500	440	280

Figure1.Job Processing Time.



Note: Graph numbered 1 shows the behavior of FCFS Algorithm whereas Line segment number 3 shows the performance of Ant Colony optimization Algorithm. The deviation shows that it consumes comparatively lesser time than that of FCFS because of grouping strategy applied on it. The disadvantage with the algorithm is it has not taken care of other parameters such as bandwidth, etc. Graph numbered 2 still shows better performance than that of the previous two because of the application both grouping and priority in the jobs in the group. The sum of all priorities of the group is accounted as priority of the group. One problem associated with this algorithm is how to handle the tie of two or more groups having same priority. This conflict can be overcome by introducing priority resolver. Though it is effective, it takes account of only jobs of similar nature. Line number 4 shows the behavior of grouping based dynamic job scheduling algorithm. Performance of grouping based job scheduling algorithm consumes lesser processing time in comparison with the all other algorithms accounted for because of the following facts.

Case 1: MIPS of job is much lesser than the MIPS of resource. The resource is not fully utilized and no further assignment is done till the MIPS of resource expires.

Case 2: MIPS of job equals to the MIPS of resource. Here the resource is fully utilized.

Case 3: MIPS of job is greater than the MIPS of resource. This assignment will not work because of the lesser MIPS of the resource. A couple of strategies is suggested. The first one is to discard the resource and wait in the queue till the job finds a suitable resource in terms of MIPS and carry out scheduling accordingly. The second option is to carry out the process partly and the part of unfinished process can be assigned to another resource subsequently.

5. CONCLUSION

In order to utilize grid resources efficiently, an adaptive fine grained job scheduling algorithm is proposed. The proposed Scheduling Model in Grid Computing is a grouping based job scheduling strategy that has taken memory constraint of individual jobs together with expected execution time at the job level into account rather than at the group level. The grouping algorithm improves the processing time of fine grained jobs. Experimental result demonstrates efficiency and effectiveness of the proposed algorithm. Though the proposed algorithm can reduce the execution time, its time complexity is high and some improvement should be done in this aspect. The proposed model reduces the waiting time of the grouped jobs. To further test and improve the algorithm, some dynamic factors such as high priority, network delay and QoS constraints can be taken into account. Advantages of this algorithm compared with others are: It reduces the total processing time of jobs. It maximizes the utilization of the resource. Minimizing the wastage of CPU power. Grouping the jobs fine-grained into grouping coarse grained will reduce the network latencies.

REFERENCES

[1] Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure.Morgan Kaufmann (1998)

[2] R.Buyya and M.Murshed, "Gridsim: a toolkit for the modeling and simulation of distributed resource management and scheduling for grid computing," Concurrency and Computation: Practice and Experience, vol. 14, 2002,pp. 1175–1220.

[3] N. Muthuvelu, Junyan Liu, N.L.Soe, S.venugopal, A.Sulistio, and R.Buyya "A dynamic job grouping-based scheduling for deploying applications with fine-grained tasks on global grids," in Proc of Australasian workshop on grid computing, vol. 4, 2005,pp. 41–48.

[4] Ng Wai Keat, Ang Tan Fong, "Scheduling Framework For Bandwidth-Aware Job Grouping-Based Scheduling In Grid Computing", Malaysian Journal of Computer Science, vol.19, No. 2, 2006,pp. 117-126 .

[5] T.F. Ang, W.K. Ng, "A Bandwidth-Aware Job Scheduling Based Scheduling on Grid Computing", Asian Network for Scientific Information, vol. 8, No. 3, pp. 372-277, 2009.

[6] V. Korkhov, T. Moscicki, and V.Krzhizhanovskaya, "Dynamic workload balancing of parallel applications with user-level scheduling on the grid," Future Generation Computer Systems, vol.25, January 2009, pp.28-34,

[7] F. Dong and S. G. Akl, "Scheduling algorithm for grid computing: state of the art and open problems," Technical Report of the Open Issues in Grid Scheduling Workshop, School of Computing, University Kingston, Ontario, January 2006.

[8] Quan Liu, Yeqing Liao, "Grouping-based Fine-grained Job Scheduling in Grid Computing", IEEE First International Workshop on Educational technology And Computer Science,vol.1, 2009, pp. 556-559.

[9] Dr. G. Sudha Sadasivam, "An Efficient Approach to Task Scheduling in Computational Grids", International Journal of Computer Science and Application, vol. 6, No. 1, 2009, pp. 53-69.

[10] K.Somasundaram, S.Radhakrishnan, "Node Allocation In Grid Computing Using Optimal Resource Constraint (ORC) Scheduling",IJCSNS International Journal of Computer Science and Network Security, vol.8 No.6, June 2008.

[11] C. Liu, and S. Baskiyar, "A general distributed scalable grid scheduler for independent tasks," J. Parallel and Distributed Computing, vol. 69, no. 3, 2009 , pp. 307-314 .

[12] Nikolaos D. Doulamis, Emmanouel A. Varvarigos ," Fair Scheduling Algorithms in Grids" IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 18, NO. 11, NOVEMBER 2007,pp. 1630- 1648.

[13] Y. C. Liang and A. E. Smith, "An ant colony optimization algorithm for the redundancy allocation problem (RAP)," IEEE Trans. Reliability,vol. 53, no. 3,2004, pp. 417–423.

[14] Vishnu Kant Soni, Raksha Sharma, Manoj Kumar Mishra ," An Analysis of Various Job Scheduling Strategies in Grid Computing **" ,** 2nd International Conference on Signal Processing Systems (ICSPS),2010 , pp.162-166.

Operating System Performance Analyzer for Low-End Embedded Systems

Shahzada Khayyam Nisar[†], Maqsood Ahmed[††], Huma Ayub[†], and Iram Baig[††]

[†]Department of Software Engineering, University of Engineering & Technology, Taxila 47050 –Pakistan
[††] Department of Computer Engineering, University of Engineering & Technology, Taxila 47050 –Pakistan

Abstract: RTOS provides a number of services to an embedded system designs such as case management, memory management, and Resource Management to build a program.
Choosing the best OS for an embedded system is based on the available OS for system designers and their previous knowledge and experience. This can cause an imbalance between the OS and embedded systems.
RTOS performance analysis is critical in the design and integration of embedded software to ensure that limits the application meet at runtime. To select an appropriate operating system for an embedded system for a particular application, the OS services to be analyzed. These OS services are identified by parameters to establish performance metrics. Performance Metrics selected include context switching, Preemption time and interrupt latency. Performance Metrics are analyzed to choose the right OS for an embedded system for a particular application.

Key Terms: *Embedded Systems, Metrics Number, Performance Analysis, RTOS.*

1. Real-time Operating Systems

An operating system is said to be real time when it schedules the execution of programs in time, handles system resources and gives a reliable basis for the development of software code. [1][2]

1.1 Components of RTOS

Most RTOS kernels consist of the following components:

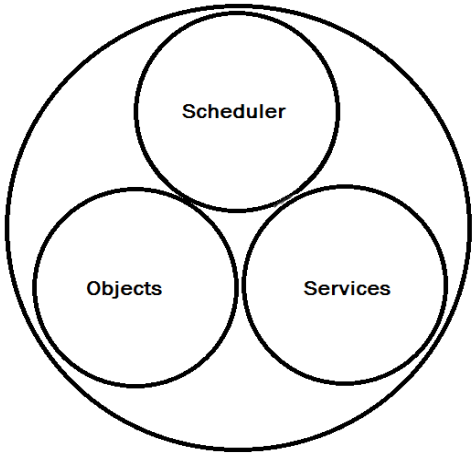


Figure 1: The normal component of the RTOS

i. Scheduler
ii. Objects
iii. Services

1.1.1 Scheduler

Scheduler is at the center of each kernel. A scheduler allows algorithms that are needed to determine what role do when.

1.1.2 Objects

The most common RTOS kernel objects are:
• *Information* --- is simultaneous and independent threads of execution that can compete for CPU execution time.
• *Semaphores* --- is a token-like object that can be raised or charged by information for synchronization or mutual exclusion.
• *Message Queues* --- are buffers that data structures that can be used, mutual exclusion, synchronization and communication by sending messages between tasks. [3]

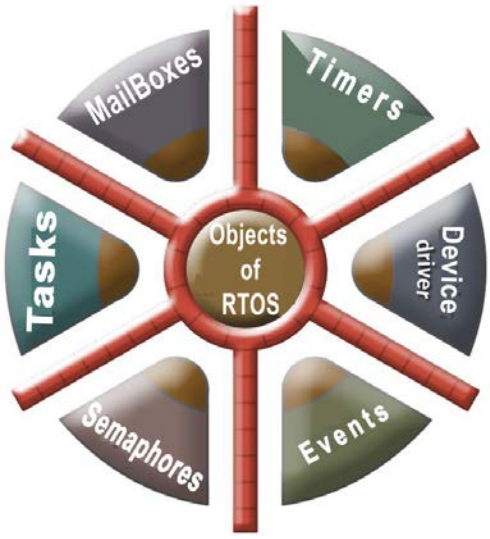


Figure 2: The Objects of the RTOS

1.1.3 Services

Most kernels provide services to assist developers for creation of real-time embedded applications. These services comprise of API calls that can be used to perform

operations on kernel objects and can be used in general to facilitate the following services:

- *Timer Management*
- *Interrupt Handling*
- *Device I / O*
- *Memory Management*

Embedded systems are used for different applications. These applications can be proactive or reactive, depends on the interface requirements, scalability, connectivity, etc. Selecting OS for an embedded system is based on an analysis of the operating system itself and the requirements of the application. [4]

2. Embedded Systems:

Embedded systems for a particular purpose are strictly monitored by the device consists of its inclusion. Embedded systems have specific requirements and pre-defined tasks unlike general purpose personal computers.

Embedded systems are programmed hardware devices. A programmable hardware chip the 'raw material' is programmed for specific applications. This is understood in comparison to older systems with hardware or systems fully functional hardware and general purpose software loaded externally. Embedded systems are a combination of hardware and software that facilitates the mass production and variety of applications. [5]

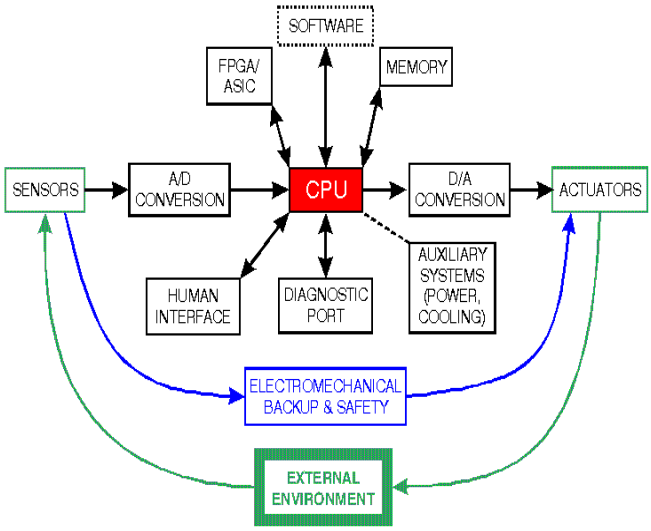


Figure 3: Schematic Embedded System

3. Selected Performance Measures for RTOS

There is a set of performance parameters that are used to analyze an operating system.

In this research, Performance Metrics consists of the following features:

i. Context Switching
ii. Preemption Time
iii. Interrupt Latency

3.1 Context Switching

It is the average time the system takes to switch between two independent active (i.e. not suspended) tasks of equal priority. Task switching is synchronous and non-anticipatory implements real-time control software for some time for slice algorithm multiplexing equal priority tasks. [6]

Task switching is fundamental performance measure of a multitasking system. Measurement attempts to assess the efficiency. The executive manipulates data structures in saving and restoring context. Data exchange is also affected by the host CPU architecture, instructions and functions.

In addition, the task is to change a measure of the manager's competence list management, as an executive normally organize their data structures in ordered lists and mixes nodes depending on the circumstances.

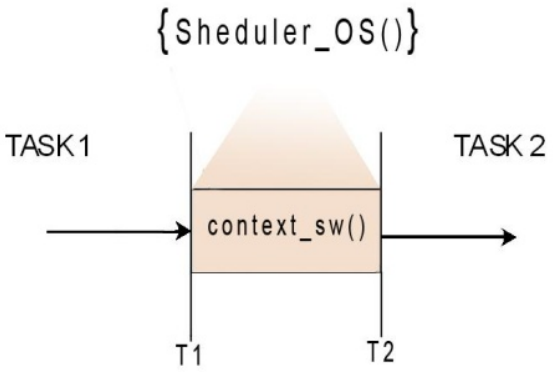


Figure 4: Context Switch Time

3.2 Preemption Time

It is the average time for a task of higher priority to wrest control of the system produces a running task a lower priority. Preemption usually occurs when the higher priority task is related to a sleep mode to a ready state in response to some external event such as when a connected device generates an interrupt, the ISR effort to wake up to the task to service the request. Preemption Time is the average time it takes the President to recognize an external event and switch control of the system produces a running lower priority task to an idle task with higher priority. [7]

Although conceptually similar to the task switch, takes first refusal usually longer. This is because the executive must first recognize waking measures and assess the relative priorities started and asked details and only then change position if necessary. Virtually all multi-use / multitasking executives assign task priorities and many allow the program designer priorities change dynamically. For this reason, together with the interrupt latency preemption is the most important real-time performance parameter. [9]

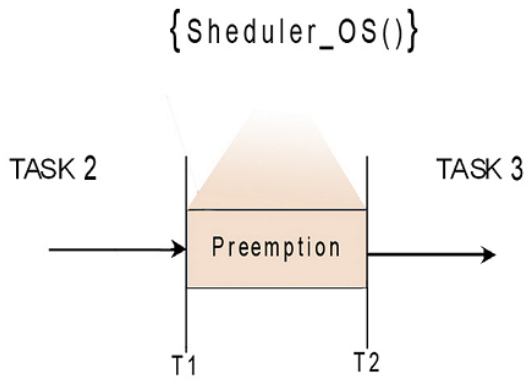


Figure 5: Preemption Time

3.3 Interrupt Latency

It is the time between the CPU receiving an interrupt request and the implementation of the first instruction in interrupt service routine. Interrupt latency reflects only the delay introduced by the executive and the processor and does not include delays on the bus or external devices. [8]

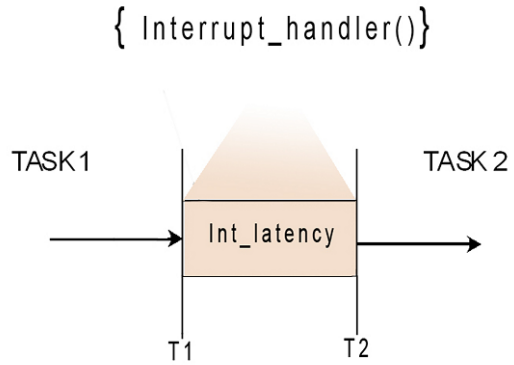


Figure 6: Interrupt Latency

4. Aims & Objectives

To choose the right operating system for an embedded system selected performance metrics are analyzed. The performance parameters are actually related to the services provided by the operating system. Improving the services provided is better operating system. In this research a Metrics Number is generated for the grade of the operating system by measuring time required for each service to occur and the number of times the service units in a complete loop. That Metrics code helps to choose the right operating system for an embedded system for a particular application.

5. Research Platform

To identify and analyze performance parameters, the environment of software and hardware is created. Three RTOS for embedded systems have been selected.

(i) SALVO
(ii) PICOS18
(iii) FREERTOS

These RTOS have free version available and also used same compiler, simulator, language and hardware platform.

6. How one RTOS differs from the other?

(i) RTOS' differ in main architechure.

(ii) Type of scheduling algorithm used in it. (Pre-emptive scheduling or co-operative scheduling).

(iii) Number of instructions of kernel without any task written to it formed after compilation of complete code. It will ultimately occupy space in ROM and RAM, so it effects memory and execution speed .

(iv) Number of tasks it can run without degrading the performace like response time. [9]

(v) Performance metrices that we have choosen i.e. Context switching Time, Preemption time and Interrupt Latency.

Applications are generated to perform multitasking. The application is analyzed in real time and monitored to extract the desired results.

In testing the performance of RTOS on 8 bit microcontroller we choose the most commonly used microcontroller family, microchip PIC 18Fxxxx class. And created a scenario in which maximum number of hardware module connected with it. The PIC18 family is being used in tremendous marketable products. So in order to check its efficiency in managing the modules controlled through the RTOS a comparison is done between different RTOS running on the same platform i.e. PIC18f452 / PIC18F4620 microcontroller, hardware modules and MPLAB simulator and PIC-C18 Compiler [10].

These modules consisting of Temperature sensor, Real Time CLOCK, UART Communication, LCD and Keypad user interfaces. This allows the 8 bit microcontroller to test the effective utilization of these resources mostly when modules are used in parallel under RTOS.

Real Time Clock and keypad working on interrupt may have same or different priorities. As this PIC range can support up to two level of interrupt handling. TIME is updated on LCD after each second through interrupt and displayed on LCD. On keypad when key is pressed interrupt is generated; shows button is pressed on LCD. Temperature sensor module and SERIAL UART running in parallel displaying data on LCD.

On hardware level the notable parameters that affect the working of RTOS in handling different modules are as follow:

7. Microcontroller

(i) Processor Clock Speed determines execution speed of RTOS.
(ii) Amount of ROM available specify the number of instructions can be stored including RTOS and task instructions.
(iii) RAM size allows the number of processes that can be run for given RTOS. [11]

(iv) STACK size provides the process local variable to accommodate.

(v) Timer used in interrupt handling and in calculating task execution time during multitasking. Timer depends upon the clock speed. [12]

(vi) That microcontroller which has no cache has RTOS architecture which is free from cache at design time.

8. Compiler and Coding Benefits

(i) For efficiency most tasks to be written in assembly.

(ii) If C language is used then optimized then good and efficient compiler that maps the C language to assembly in minimum no. of instruction is the ultimate goal.

(iii) Optimized coding technique used in defining tasks. [13]

9. Hardware Information

(i) PIC18F452 has Harvard architecture (has a separate instruction and data bus).

(ii) 40 pin IC.

(iii) 1536 BYTES on chip RAM.

(iv) 32KBYTES FLASH memory for Program storage.

(v) Maximum 16384 single word instructions can be placed in FLASH memory.

(vi) 2 interrupt priority levels.

(vii) 8MHZ internal oscillator or up to 20 MHz external oscillator can be used.

Block Diagram

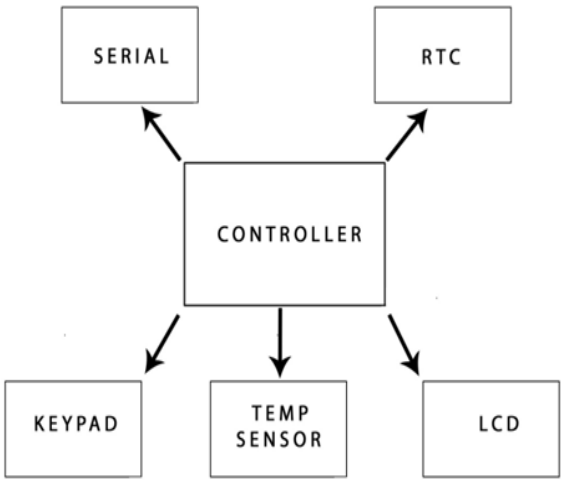


Figure 7: Block Diagram of Hardware

10. Analysis

As the application comprises of multi tasking and interrupts therefore, the Performance Metrics selected comprises of three performance parameters. These are

1. *Context Switch time*
2. *Preemption time*
3. *Interrupt Latency*

For measuring the Metrics Number there are two requirements

1. *Time "T" for each parameter*
2. *Number of times "W" the parameter is called*

11. Calculation of Time for Each Parameter

In order to calculate the time required for each parameter it is necessary to identify where these parameters are called. To calculate the time hardware timers are used. Break points are given to the entry and exit of a parameter. Timer is initialized on the entry break point and terminated on the exit. The time calculated gives us the time for a parameter. Same procedure is followed for rest of parameters.

For Example: The Context Switch Time is calculated by marking Breakpoints on the start and end of the context switch service of the operating system. Hardware Timer is used for calculating the time. Timer is initialized at the start break point and terminated at the exit break point. This will give us the time required by the operating system for context switch for two tasks.

12. Calculation of Weight of Each Parameter

To find out the number of times the parameter is called a variable is defined in each parameter. For each time the parameter is called that variable is incremented. The incremented value is displayed on the Display screen. For each parameter that variable is defined and measured for one complete iteration of the application.

RTOS was made working in this WAY For testing of performance:

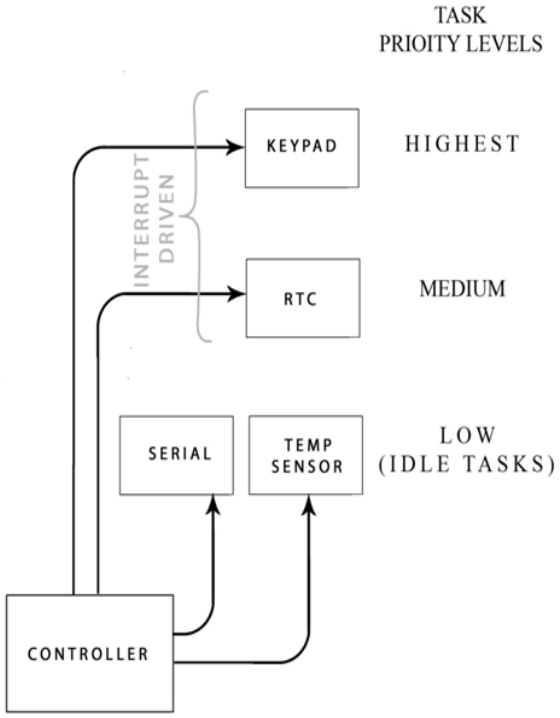


Figure 8: RTOS Scheduling Policy

To find the time necessary to provide the required service timers are initiated at the beginning and end when the loop

is complete. This will allow time an operating system to provide the service that is their own ability. A variable is initialized for each service for which time is expected to learn that the number of times that the service has been called. This will give the following values:

Time T are the observed values of Performance Parameters in microseconds. Number of times the performance parameter occur is given by the **Weight W,** the product of T and W is given by **TxW** and total sum of all the three product is given by **ΣTxW** in microsecond.

13. Metrics Number

For understanding we did comparison by analyzing the real functionalities / usage of context switching, preemption and interrupt latency in applications and compared accordingly.

14. Generation of Metrics Number for Free RTOS

- ***Context Switching*** is measured by using yield() functions. It means to give control to other task when both tasks have the same priorities. Execution time of task yield() has to be considered because it is the part of kernel or scheduler.

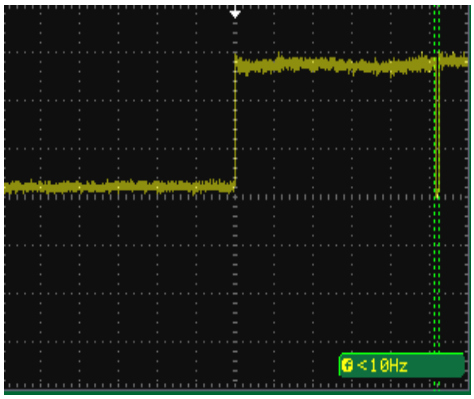


Figure 9: Context switching for Free RTOS

- ***PreemptionTtime*** is measured when task priority change or giving execution time to higher priority task by changing the task priority to higher level. we also include the time of maketask_priorityhigh() function.

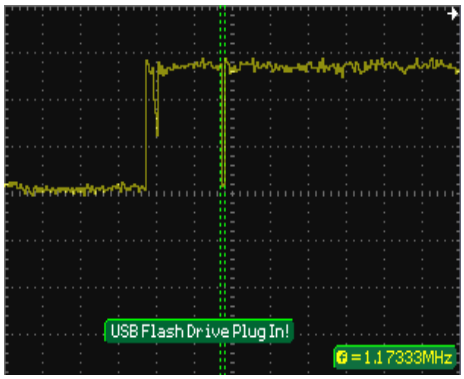


Figure 10: Preemption Time for Free RTOS

- ***Interrupt Latency*** is the measured time between when external interrupt came and ISR related to that interrupt start executing.

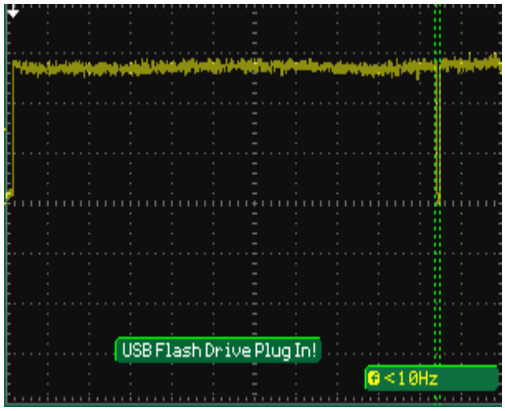


Figure 11: Interrupt Latency for Free RTOS

Table 1: Generation of Metrics Number for FREE RTOS

Performance Parameter	Time T (µ sec)	Weight W	T X W	ΣTxW (µ sec)	METRICS NUMBER 1/ ΣTxW
Context Switching Time	T_{CS} =7	W_{CS} = 55	385	5359	**186.60**
Preemption Time	T_P =15	W_P = 67	1005		
Interrupt Latency	T_{IL} = 3.5	W_{IL} = 1134	3969		

Generation of Metrics Number for PICOS18

- ***Context Switching*** is measured by then each time the task 0 sends an event to the task 1 the time to switch from task 1 to task 0.

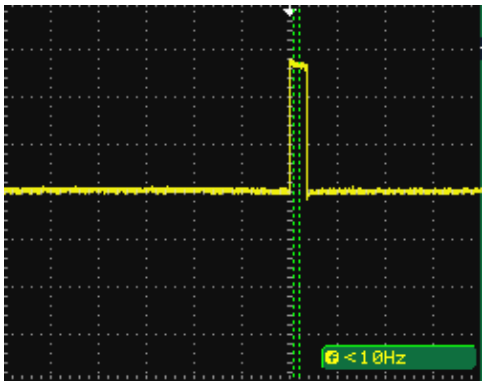


Figure 12: Context switching for PICOS18

- ***Preemption Time*** is measured when task priority change or giving execution time to higher priority task by changing the task priority to higher level.

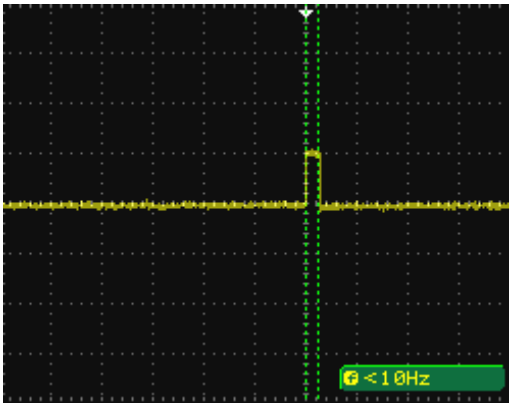


Figure 13: Preemption Time for PICOS18

- ***Interrupt Latency*** is the measured time between when external interrupt came and ISR related to that interrupt start executing.

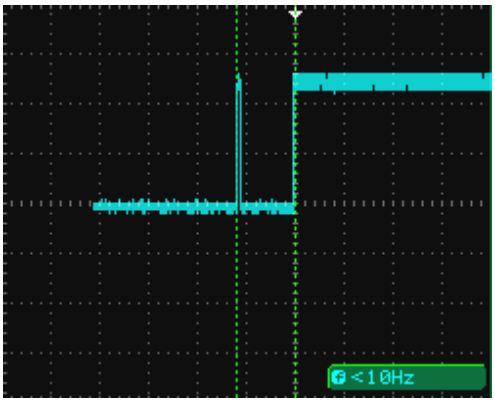


Figure 14: Interrupt Latency for PICOS18

Table 2: Generation of Metrics Number for PICOS18

Performance Parameter	Time T (µ sec)	Weight W	T X W	ΣTxW (µ sec)	METRICS NUMBER 1/ ΣTxW
Context Switching Time	TCS =47	WCS = 55	2585	60290	**16.58**
Preemption Time	TP =15	WP = 67	1005		
Interrupt Latency	TIL= 50	WIL= 1134	56700		

15. Generation of Metrics Number for SALVO RTOS

Co-operative context switching depends on the task that is currently running. The current task calls for other to switch to other for its working. But in preemptive context switching the scheduler do not take care of the running task when higher priority task occurs. SALVO is the only RTOS that is not preemptive but co-operative RTOS. There are upto 15 levels of priorities.

- ***Context switching*** is measured by Using OS_Yield() functions. Its mean giving control to other task when both task have the same priorities. Execution time of task yield() has to be considered because it's part of kernel or scheduler.

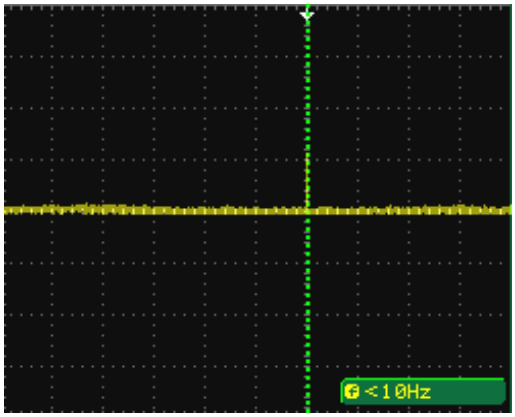


Figure 15: Context switching for SALVO RTOS

- ***Preemption time*** is measured when task priority change or giving execution time to higher priority task by changing the task priority to higher level. We also include the time of OS_SetPrio() function.

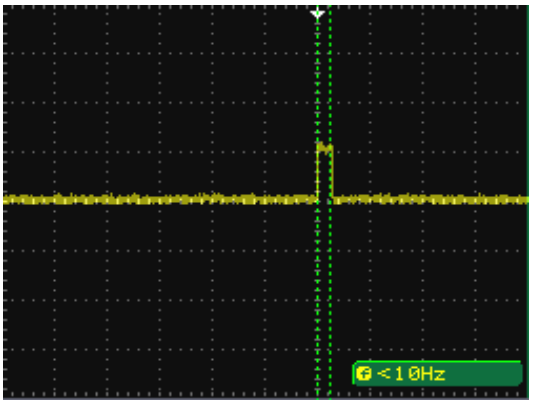


Figure 16: Preemption Time for SALVO RTOS

- ***Interrupt latency*** measured the time between when external interrupt came and ISR related to that interrupt start executing.

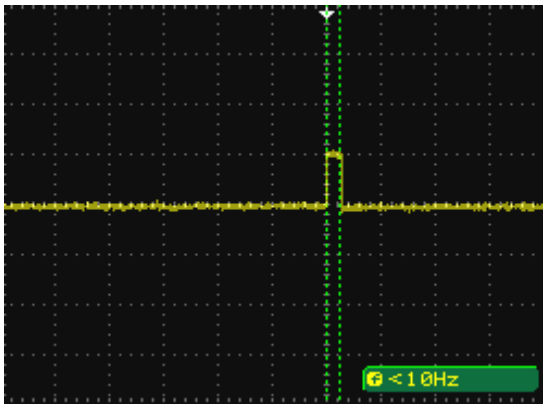


Figure 17: Interrupt Latency for SALVO RTOS

Table 3: Generation of Metrics Number for SALVO RTOS

Performance Parameter	Time T (µ sec)	Weight W	T X W	ΣTxW (µ sec)	METRICS NUMBER 1/ ΣT xW
Context Switching Time	T_{CS} =10	W_{CS} = 55	510	4149	**241.02**
Preemption Time	T_P =12	W_P = 67	804		
Interrupt Latency	T_{IL} = 2.5	W_{IL} = 1134	2835		

In Tables, Column 1 lists performance parameters, for which the Metrics numbers must be generated. In column 2, the time T required for each performance parameter in micro-seconds is listed. "T" is the measurement of time intervals between time initialization and termination for each performance parameter. Column 3, the number of times the performance parameter occurred is denoted by "W". "W" is the weight given to each performance parameter for that specific application. Column 4 is the weighted value of each performance parameter. Column 5 is the summation of all weighted measurements. The inverse of the sum of the weight measurements gives us a number Metrics. *Larger Metrics numbers are better operating system for that specific application*.

16. Conclusion

The measurement of this Metrics Number has a great significance in selection of right operating system for a specific application. If another operating system is selected and same application is used with same hardware then the weights will remain same however the time observed for each performance parameter will be different. If the resulting Metrics number is greater, then this operating system is best for that environment. The Metrics Number generated will help us in rating the operating system. This will help us in deducing a procedure for selecting the right Performance Metrics. Having right performance metrics will help us to calculate metrics number which will help us in selecting right operating system for an embedded system for a specific application.

17. Future Directions

In this paper a method to analyse performance metrics of operating system in real-time embedded systems is described. For future work it is recommended that if the methodology for the application processing time is formulated then right processor can also be selected for the embedded system. This will help the designer to make an efficient embedded system with a right Real Time Operating System.

References

[1] Wei-Tsun Sun; Zoran Salcic; , "Modeling RTOS for Reactive Embedded Systems," *VLSI Design, 2007. Held jointly with 6th International Conference on Embedded Systems., 20th International Conference on* , pp.534-539, Jan.2007 doi:10.1109/VLSID.2007.111

[2] Su-Lim Tan; Tran Nguyen Bao Anh; , "Real-time operating system (RTOS) for small (16-bit) microcontroller," *Consumer Electronics, 2009. ISCE '09. IEEE 13th International Symposium on* , pp.1007-1011,25-28 May 2009 doi: 10.1109/ISCE.2009.5156833

[3] Baynes, K.; Collins, C.; Fiterman, E.; Brinda Ganesh; Kohout, P.; Smit, C.; Zhang, T.; Jacob, B.; , "The performance and energy consumption of embedded real-time operating systems," *Computers, IEEE Transactions on* , vol.52, no.11, pp. 1454- 1469, Nov. 2003 doi: 10.1109/TC.2003.1244943

[4] Hessel, F.; da Rosa, V.M.; Reis, I.M.; Planner, R.; Marcon, C.A.M.; Susin, A.A.; , "Abstract RTOS modeling for embedded systems," *Rapid System Prototyping, 2004. Proceedings. 15th IEEE International Workshop*, pp. 210- 216, 28-30 June 2004 doi:10.1109/IWRSP.2004.1311119

[5] He, Z.; Mok, A.; Peng, C.; , "Timed RTOS modeling for embedded system design," *Real Time and Embedded Technology and Applications Symposium, 2005. RTAS 2005. 11th IEEE* , pp. 448- 457, 7-10 March 2005 doi:10.1109/RTAS.2005.52

[6] Suk-Hyun Seo; Sang-won Lee; Sung-Ho Hwang; Jae Wook Jeon; , "Analysis of Task Switching Time of ECU Embedded System ported to OSEK(RTOS),"*SICE-ICASE,2006. International Joint Conference* , pp. 545-549, 18-21 Oct. 2006 doi: 10.1109/SICE.2006.315544

[7] Kavi, Krishna; Akl, Robert; Hurson, Ali; "Real-Time Systems: An Introduction and the State-of-the-Art" *John Wiley & Sons, Inc. Wiley Encyclopedia of Computer Science and Engineering;* SN: 9780470050118; 2007; doi: 10.1002/9780470050118.ecse344

[8] El-Haik, Basem; Shaout, Adnan; "Design Process of Real-Time Operating Systems (RTOS)" *John Wiley & Sons, Inc. Software Design for Six Sigma;* pp. 56-76; SN: 9780470877845; 2010; doi: 10.1002/9780470877845.ch3

[9] Edwards, Stephen A.; "Real-Time Embedded Software"; *John Wiley & Sons, Inc.; Wiley Encyclopedia of Electrical and*

Electronics Engineering; SN: 9780471346081; 2001; doi: 10.1002/047134608X.W8113

[10] Weiss, K.; Steckstor, T.; Rosenstiel, W.; , "Performance analysis of a RTOS by emulation of an embedded system ," *Rapid System Prototyping, 1999. IEEE International Workshop on* , pp.146-151, Jul 1999 doi:10.1109/IWRSP.1999.779045

[11] Stepner, D.; Rajan, N.; Hui, D.; , "Embedded application design using a real-time OS," *Design Automation Conference, 1999. Proceedings. 36th* , pp. 151-156, 1999 doi:10.1109/DAC.1999.781301

[12] Elsir, M.T.; Sebastian, P.; Yap, V.V.; , "A RTOS for educational purposes," *Intelligent and Advanced Systems (ICIAS), 2010 International Conference on* , pp.1-4, 15-17 June 2010 doi:10.1109/ICIAS.2010.5716166

[13] Cena G., Cesarato R., Bertolotti I.C. "An RTOS-based design for inexpensive distributed embedded system" (2010) *IEEE International Symposium on Industrial Electronics*, art. no. 5636340, pp. 1716-1721.

Transmission System Planning in Competitive and Restructured Environment using Artificial Intelligence

Engr. Badar Ul Islam
Assistant Professor, Department of Electrical Engineering
NFC-Institute of Engineering & Fertilizer Research, Faisalabad – Pakistan.

Engr. Syed Amjad Ahmed
Associate Professor, Department of Mechanical Engineering
NFC-Institute of Engineering & Fertilizer Research, Faisalabad – Pakistan.

ABSTRACT

This paper picturesquely depicts the changing trends and values under new circumstances which are developed in electric power system i.e. generation side and partly on the way in transmission and distribution network. A very clear advocacy about the changing trends from vertical integrated setup to the horizontal disintegrated setup is explained in very simple way. All utilities are passing through the phase of disintegration globally; it is obvious that the same effect is also putting impression on the electrical power sector. This effect is designated as restructuring and competitive environment of public utilities. This specific approach means, that the public now demands to break the monopolistic approach and wants that the public utilities must be operated by public themselves but under umbrella of some regulatory body who can watch their interests and legislate rules which helps the masses in getting the better service than that they have. A clear comparison is also presented between the past/existing standard practices with the future methodology of transmission system planning. It also suggests that necessary analysis may also be done on computer by using different models and with the use of artificial intelligence and the expert system is considered to be the best with its features for transmission system planning.

Key words:

Restructuring, Congestion, (n-1) configuration, load forecasting, algorithm, expert system (ES).

1. INTRODUCTION

The World has changed its pace and developed into a global village, keeping close all the different sectors.

Electrical power being such an important requirement of all the sectors and requires to be developed, from its generation sector to the distribution end, that it should be developed in a more executive fashion then at present. A lot of work has already been done in the field of generation and the distribution sector, but the transmission sector still lags behind than the later. This paper specifically focuses about the transmission system planning, with an open explanation to the problems commonly arises during transmission planning by the transmission planners.

It is important to mention that in the past and also at present, continuous efforts are made to improve and develop the electrical transmission system which is more efficient, reliable and cost effective. A number of different techniques were used.

At present, latest computerized techniques are used. The Artificial Intelligence (AI) helps to develop the new transmission plans in more versatile manner with elaborated picture and with more options.

The World has change its attitude and the trend

is now of commercialization and this commercialization develops the approach of a deregulated environment in a competitive frame of reference. The power system mainly consists of generation, transmission and distribution. Generation and the distribution system is now playing its role both in the public and private sector but uptill now, the transmission sector is only working in a public sector or to some extent under corporative culture. This does not fulfill the requirements as that of a deregulated and competitive environment requires. [1]

This indication clearly opens the vistas to explore the new means of transmission system planning using artificial intelligence, by any of its application which is more beneficial and covers all the aspects as desired by the humanity.

In order, to manage and distribute electrical power in an effective and economical way, a properly planned transmission system is the basic and key requirement. Serving as a back bone between the generation and distribution end, the transmission system planning must be done in such a way to accommodate all the important aspects which are required to supply the power in an efficient, reliable and cost effective manner. The following are the main features which needs due consideration while doing the transmission system planning:

- Transmission system planning in accordance with the forecasted load.
- Segregation of forecasted load during peak hours and off peak hours.
- Suitable room to accommodate all the power producers.
- Provision for (n-1) link in transmission network for all type of voltage levels.
- Substation/grid stations of sufficient capacity to cater the generated load.
- Alternate arrangement in transmission network, while when one is under maintenance/fault.
- Provision to accommodate 50% of the transmission line load when one line is overloaded.
- Congestion management.
- Always have the sufficient capacity in accordance with the time frame of load required.

2. STANDARD PLANNING TECHNIQUES

The Figure – 1 describes the different phases required for transmission system planning. This flow chart is a basic one and all the stages involve in this are equally applicable in the past, even at present and also fulfills the basic requirements for future transmission system planning [1].

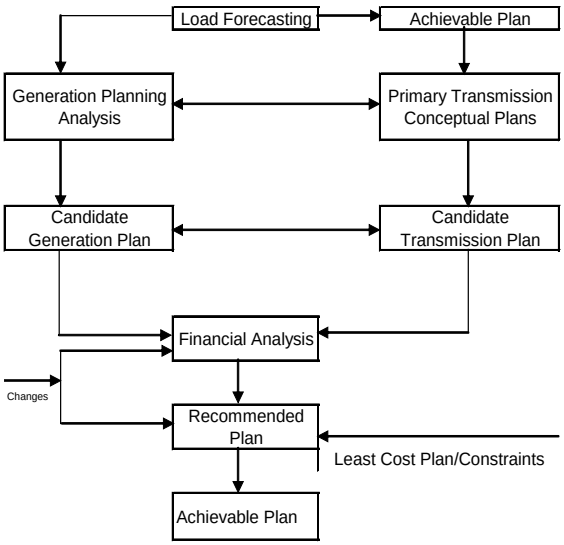


Figure: 1 Standard Practice of Transmission System Planning.

3. WHY THERE IS A D EMAND FOR DEREGULATION?

Importance of public utilities is evident right from the day, the mankind came into existence. As the generation grows and population increases, needs also increased and human beings are continuously trying to find better solutions and ways to cope-up with these. Amongst different public utilities are water, natural gas, communication network, electric power etc. Every utility has its own importance, but the electric power is at the top in the present era, right starting from its generation level to the distribution level. It is obvious that in this global era of fast track development, all the public utilities cannot be developed fully by the state due to financial constraints .Especially under-developed countries cannot take this sector fully [1].

Keeping in view, the above picture, now-a-days, there is change in all sectors including power sector globally. This global village now demands change in power sector and expects more than is available and it can only be achieved by applying different modern techniques and devise new methodologies to achieve the desired goals. In this current situation, the actual depiction of image of power sector in restructured environment demand new methodology with new principles for this deregulated, restructured and competitive environment. Further, the

approach must be rational but not limited to one state [1].

This thought of deregulation has given birth to change the system from vertically integrated system to convert it into horizontal one i.e. breaking up the monopolistic approach and with new system every one is free to come into power sector market and proves that, the services provided by his set-up is better than other, a more competitive approach in deregulated and restructured, competitive environment.

In restructured environment, especially when the services are changing mode from vertical integrated setup to a horizontal disintegrated restructured setup, there is an utmost need to have one regulatory body which regulates all the matters of power sector in all respects. [5].

4. ROLE OF REGULATORY BODY

The regulatory body is required for continuous monitoring all the activities which are going in the electrical sector and different reforms are to be formulated for the betterment of the public, like formulation of new principles, procedures, legislation, practices etc. of the electrical sector of any country. [5].

5. FUTURE METHODOLOGY OF TRANSMISSION SYSTEM PLANNING

The Figure - 2 is self explanatory and explaining the different steps involved in transmission system planning for the future. This flow chart also covers the practice followed in the past and shown as in figure-1 above and in addition to that different important aspects for which the specific care is needed are also added which gives a more better, reliable and cost effective service as compare to the past/existing standard practices [7].

See Figure : 2 as Annexure

6. SIGNIFICANCE OF AI

AI plays a pivotal role in every field existing in the world and especially in the filed of sciences and in particular in engineering sector. Its significance is evident from the fact that AI attempts to understand intelligent entities. The intelligent entities are interesting and useful. AI has produced many significant and impressive products. It is clear that computers with human level intelligence would have a huge impact on our every day lives and on future course of civilization.

AI helps in understanding how to see, learn, remember and reason could be done. In this, the computer provides a tool for testing theories of intelligence. [4]

7. DIFFERENT FIELDS OF ARTIFICIAL INTELLIGENCE

The following are the types of AI:-

i. Expert System
ii. Fuzzy Logic
iii. Neural Network [2 , 3]

7.1 Expert System

An expert system is commonly known as knowledge based system, based on computer program that contains the knowledge and analytical skills of one or more human experts related to a specific subject. [2]

The most common form of an expert system is a computer program with the set of rules that analyzes information about a specific class of problems and recommends one or more courses of user action. The expert system also provides mathematical analysis of the problem. It utilizes the reasoning capability to reach on the conclusions. [3]

7.2 Fuzzy Logic

Computers are too logical and they only deal in true or false, yes or no etc.

Fuzzy logic allows a computer to deal in every day human language and actually process terms such as probably, unlikely, quite, near etc. Such terms can take their place in computations allowing the computer to arrive at verifiable results from fuzzy inputs. The logic used is mathematically verifiable so results for the process can be trusted.

Fuzzy logic is derived from fuzzy set theory dealing with reasoning that is approximate rather than precisely deduced from classical predicate logic. It can be thought of as the application side of fuzzy set theory dealing with well thought out real world expert values for a complex problem.

7.3 Neural Network

Traditionally, the term neural network had been used to refer to a network or circuit of biological neurons. The modern usage of the term often refers to artificial neural networks, which are composed of artificial neurons or nodes. Thus the term 'Neural Network' has two distinct usages. [9]

- Biological neural networks are made up of real biological neurons that are connected or functionally related in the peripheral nervous system or the central nervous system. In the field of neuroscience, they are often identified as groups of neurons that perform a specific physiological function in laboratory analysis.

- Artificial neural networks are made up of interconnecting artificial neurons (programming constructs that mimic the properties of biological neurons). Artificial neural networks may either be used to gain an understanding of biological neural networks, or for solving artificial intelligence problems without necessarily creating a model of a real biological system.

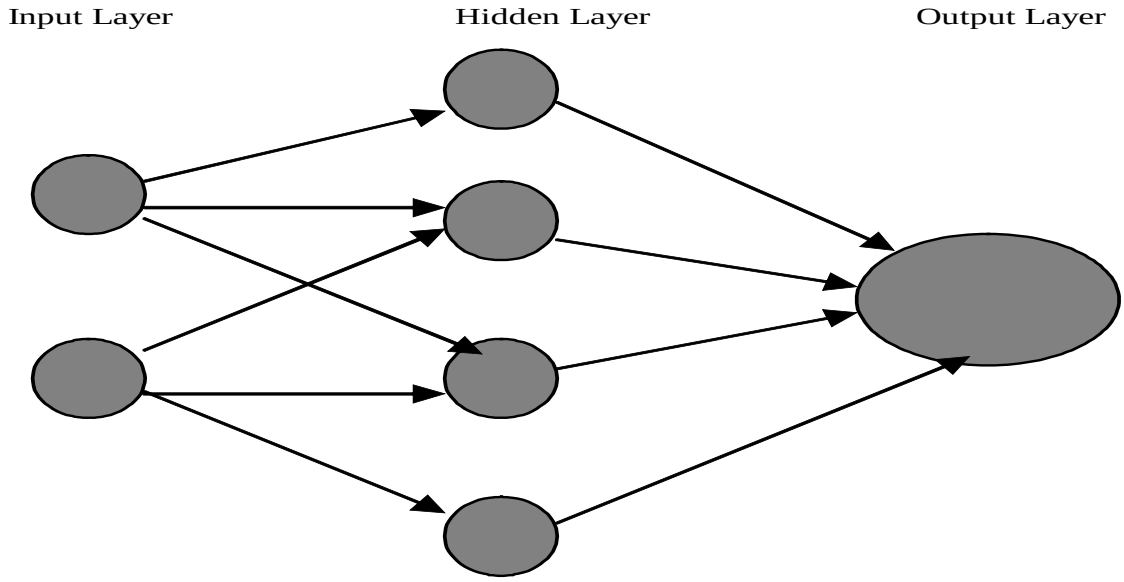


Figure – 3 Simplified view of an artificial Neural Network

8. ELECTRICAL TRANSMISSION NETWORK WITH DIFFERENT VOLTAGE LEVELS.

In any country, the electrical transmission network is always consists of different voltage levels, according to the needs of the area linked with. This electrical network also shows the location of generating stations at different point's along with the generating capacities. This addition of power changes the voltage profile of the electrical network and stables the voltage level. This addition of power also improves the power factor of the system.

The Figure – 4 is a part of complex electrical network with different voltage levels i.e. 220KV and 132KV.

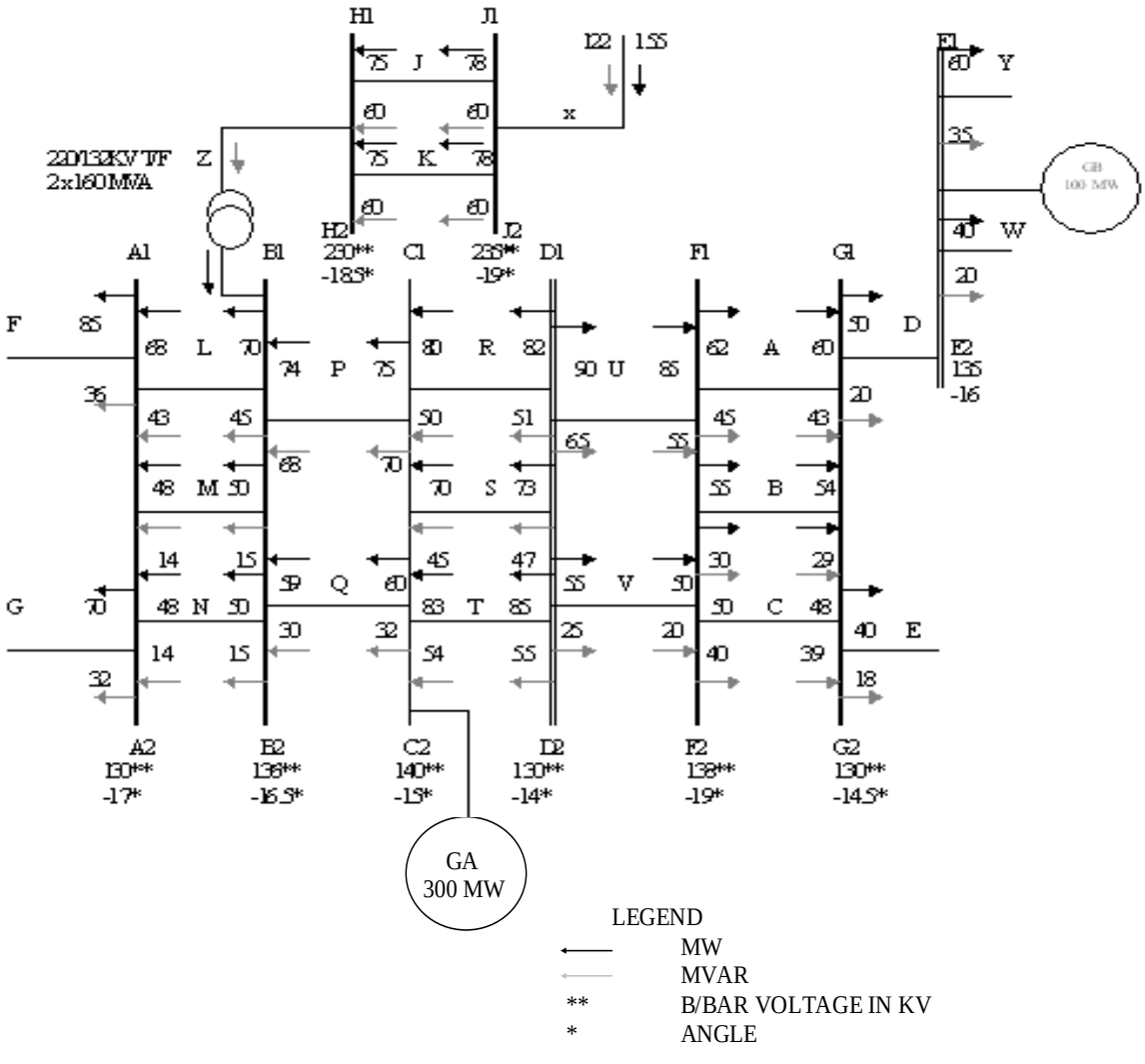


Figure 4:- A part of 220 KV and 132KV Complex Electrical Transmission System

9. TRANSMISSION SYSTEM PLANNING IN TRADITIONAL SETUP VS RESTRUCTURED ENVIRONMENT.

A comprehensive comparison of the transmission system planning in traditional setup and deregulated setup is given under: -

Item	Traditional Setup	Deregulated Setup
Load forecasting	This is done but not on the actual basis, without having a in depth look of economic growth and other parameters	Of immense importance and to be done on actual basis.
Segregation of load	Totally neglected	For accurate transmission planning and to create the hedge, segregation of load is required between peak and off peak hours
Availability of transmission line	Independent power producers (IPP's) were invited but the system lacks to accommodate the generated power	Key feature, in order to develop a reliable and cost effective transmission system, suitable room must be provided before

n-1 link	Already n-1 links are available but at certain points, this provision is not available	This must be given to all the links, irrespective of load requirement
Substations/ grid stations	These must be augmented according to the load requirements in order to provide reliable and stable electrical power	Giving equal importance as that of transmission system planning, if augmentation or construction of new grid station are required, these must be planned in parallel to the transmission planning
Alternate arrangement and to accommodate 50% of the required load	Proper alternate arrangement is not available to accommodate the 50% of the required load during maintenance or overloading	During transmission planning, the network should be designed as having the capacity to accommodate at least 50% of the load requirement
Congestion management	Totally neglected	Key feature during the transmission planning and even of more important nature in deregulated setup when there are number of IPP's desire to supply power on the system

10. KEY FEATURES IN TRANSMISSION SYSTEM PLANNING.

Transmission system planning is the most important sector of electrical network and has the following key aspects:

- Load forecasting
- Primary transmission conceptual plan
- Generation planning activities
- Candidate transmission plan
- Candidate generation plan
- Financial analysis
- If no constraints and the plan is technically feasible and economical viable then
- Recommended plan
- Approval from regulatory body
- Achievable plan

In both the setups i.e. traditional and deregulated, the above noted are the key features to have a proper transmission plan.

All the features have significance but the key feature in the transmission system planning is the load forecasting. This is the load forecasting which actually gives the signal in the system and then directs the attentions of the engineers to study and make necessary changes in the system according to the requirements. No transmission system can be planned without having proper forecasted figures.

It is the load forecasting which indicates that the requirement of electrical power is increased from the present load requirement and in accordance to that new-projected figures of load, different transmission plans are required to be proposed in connection with the increase of power demand, which diverts the attention to install new power plants. This study ultimately leads to have a candidate transmission plan and at the same time, the requirement of a new power plant.

After the selection of candidate transmission and generation plan, financial analyses are to be made in order to calculate the benefit cost ratio. If no constraints are there, the project is discussed with the approving authority and if the authority agrees with all the aspects, the project is called the achievable plan.

There is no specific change in the ways of transmission planning as of traditional setup with the deregulated setup except a special attention towards the environment is required as an international law.

Again coming back, showing the key importance of load forecasting, if this is done on actual grounds, the transmission system planning is automatically be the actual one.

11. TRANSMISSION CONGESTION MANAGEMENT

Transmission congestion occurs when there is

insufficient transmission capacity to simultaneously accommodate all requests for transmission service within a region. Historically, vertically integrated utilities managed this condition by constraining the economic dispatch of generators with the objective of ensuring security and reliability of their own and/or neighboring systems.

The top priority items during transmission system planning:

- Congestion management
- Cost recovery
- Market monitoring
- Transmission planning
- Business and reliability standards
- Transmission rights

11.1 Congestion zones

The zones are defined such that each generator or load within the zone has a similar effect on the loading of the transmission lines between zones. Once zones are defined, any imbalance between load and generation within a zone is assumed to have the same impact on inter-zonal congestion. Zone boundaries are reexamined annually to see if generation, load, or transmission patterns have changed enough to warrant changing the zones. The zones are designed to capture the "commercially significant constraints".

11.2 Improvements in congestion management

Effective transmission system planning also addresses to start charging customers directly for the commercially significant congestion as proposed by the regulatory body.

Deregulation and policies of open access, allocation of scarce transmission resources has become a key factor for the efficient operation of electricity markets as well as reliability and control of market power. This trend emphasizes for the re-enforcement and expansion of the transmission capacity in accordance with the demand of power and the emerging trend to transfer power over long distances. These trends are important for congestion management to structure and facilitate economically efficient allocation of transmission capacity.

12.0 OPERATIONAL POLICIES

The importance of policy, procedures etc. are definite in every field. The same is the case with the electrical power. Operational policy is required in both the cases whether the system is a regulated one or a deregulated. Globally, different type of operational polices were in practice, when the electric utility is state owned and governed by the state. Now, the trend is changing and the process of disintegration of electric utility is on its way and also the approach is now more inclined to a horizontal deregulated setup. In this regard, the important features which must be taken into account and considered while devising the transmission system plans. This operational policy helps in explaining the importance and need for measurement and standards for planning the transmission network in competitive and deregulated setup [5].

The under mentioned points must be taken into account while doing the transmission system planning.

- Increased growth in the number and complexity of transactions.
- Increased number of market players and their information needs.
- Competitive metering of energy generation—including distributed generation—and ancillary services at the supplier and customer levels.
- Monitoring bulk power flows and transactions.
- Monitoring transmission and distribution system conditions [5].
- Monitoring power quality along these systems and in customer facilities.
- Tracking/tagging of power flows to assign cost responsibility for congestion on overloaded lines and constrained interfaces.

In general, deregulation will lead to changes in several important electrical power industry characteristics:

- Services will be unbundled, and is necessary to separately evaluate each type of transaction.
- Time frames will shorten. New services will be measured over seconds and minutes instead of hours.
- Transaction sizes will shrink. Instead of

dealing only in hundreds and thousands of MW, it will be necessary to accommodate transactions of a few MW and less.

- Supply flexibility will greatly increase. Instead of services coming from a fixed fleet of generators, service provision will change dynamically among many potential suppliers as market conditions change.
- Who will pay the cost of Reactive power [5]

Finally, the different issues discussed above are also of key importance but certain other technical features that must be taken into account before developing transmission system plans in competitive and restructured environment. These are

- Increased transmission demand,
- Service quantification,
- Reliability criteria,
- Real-time electric pricing,
- Unbundling of ancillary services,
- Reduced generator and transaction sizes,
- Power quality, and
- Supplier choice.

13.0 EXPERT SYSTEM (ES)

The significance of artificial intelligence (AI) and the philosophy of AI were already discussed in preceding paragraphs. Further, different systems were also discussed and prove that ES is better then the others.

As already discussed the importance of load forecasting with reference to transmission system planning, so an expert system is designed for electrical load forecasting based on multiple linear regression.

The model for the hourly load at each of the considered time intervals has the form;

$$Yi(t) = Ai + Bi(Td(t) - Tci) + Ci(Td(t) - Tci)2 + Di(Td(t) - Tci)3 + Ei(Tp(t) - Tpi) + Fi(Tava - Tavb) + Gi(Td(t) – Td(t - 1)) + Hi(Td(t - 1) – Td(t - 2) + Ii(Td(t -2) – Td(t - 3))$$

and where

yi(t) = load at hour t in the interval of the day

AI = base load component (regression constant coefficient)

Bi through Li = regression coefficient of weather sensitive component

Td(t) = dry bulb temperature at time t, f=deg F

(which will be clamped at the cut off value if necessary)

Tp(t) = dew point temperature at time t, f=deg F

(which will be clamped at the cut off value if necessary)

Tava = average dry bulb temperature of previous 24 hours to the time t, deg F

Tavb = Tava lagged 3 hours, deg F

Tci = cut off dry bulb temperature for the interval I in the season, deg. F

Tpi = cut off dew point temperature for the interval I in the season, deg. F

v(t) = wind speed at time t, miles/hour

we have the values of above parameters estimated for different time intervals

The Figure – 5, shows the flow for the Algorithm and results are shown as in Figure – 6.

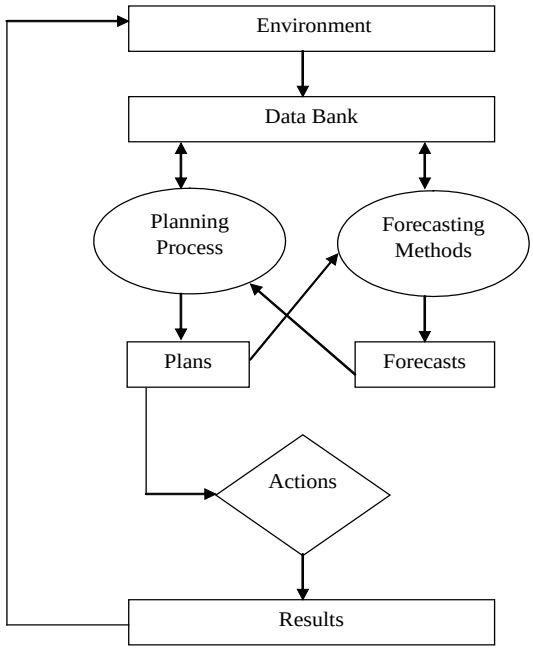


Figure – 5: Flow Diagram for Algorithm

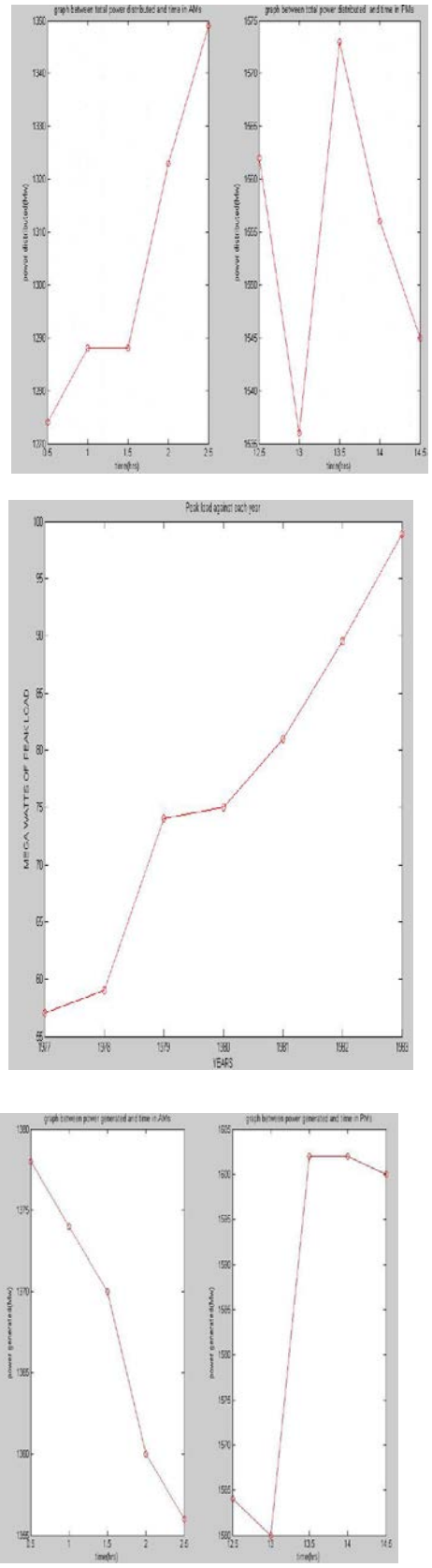


Figure 6 : Graphs

14. CONCLUSIONS

The electrical power transmission system is the backbone of the electrical network. In order to develop, an efficient, reliable and cost effective system, this needs specific attention towards the planning part. Effective planning gives better results and more reliable service to all categories of consumers, with an open, fair and free access on discriminatory basis.

Effective planning can only be done with the help of artificial intelligence and expert system. In this present modern era, the role of computerized technology is in all the fields and in order to study over wide options and to do the better planning of the existing system and to develop the new ones, expert system is the most effective solution.

The transmission system planning serves as a backbone in the electrical network system. All the parameters involved in electrical network have unique importance but in transmission system planning, the load forecasting is the primary feature to develop the stable, reliable and cost effective system. In order to obtain the better results, the use of artificial intelligence is made and an algorithm is develop in ES, so as to have better load forecasted figures which helps to plan the system in better way and in less time.

ACKNOWLEDGEMENTS

The author is thankful to the WAPDA for providing the useful information about the transmission system planning and techniques. Also grateful to the NFC-IEFR Institute, Faisalabad for providing the help and technical facilities.

REFERENCES

[1] Rajat K. Deb, Pushkar Wagle, and Rafael Emmanuel A. Macatangay,

"Generation and Transmission Investments in Restructured Electricity Markets", 2005. Available http:// www.energyonline.com.

[2] "Artificial Intelligence Structures and Strategies for Complex Problem Solving" by George F Luger, Fourth Edition, July 2001.

[3] "Understanding Artificial Intelligence" by Henry C. Mishkoff, BPB Publications, B-14, USA.

[4] "Expert System Principles and Programming" by Giarratano and Riley, Third Edition, PWS Pubications, University of Houston, Clear Lake, 1999.

[5] Aamir Mahboob Ilahi & Dr. Suhail Aftab Qureshi, Part-1, "Transmission System Planning in Competitive and Restructured Environment" paper published in IEEEP journal in 2000.

[6] Aamir Mahboob Ilahi & Dr. Suhail Aftab Qureshi, "Deregulation of Power Utility Services for Transmission System Planning" paper published in University Research Journal published by University of Engineering & Technology Lahore, Pakistan 2004.

[7] Paper published in 17th IASTED proceedings at "The International Association of Science and Technology for Development" Canada on "Current trends in transmission system planning needed in competitive and deregulated environment using Artificial Intelligence" and presentation was made on May 24th to 26th, 2006.

[8] Paper published in IEEEP Research Journal 2005 on "Application of Artificial Intelligence in Present World and its importance in the field of Engineering".

[9] "A Multi Stage Intelligent System for Unit Commitment" by Z. Quyang, S.M.Shahidehpour, Department of Electrical and Computer Engineering, Illinos, Institute of Technology, Chicago.

[10] "AI in Energy System and Power" First International, Inter Discipline Research Canada, Symposium, February 10, 2006.

Robust RSA for Digital Signature

Mr.Virendra Kumar, Mr. Puran Krishen Koul

Department of Computer Science
MTU Noida
CET-IILM-AHL
Knowledge Park-2
Greater Noida
G.B.Nagar
UP 201306,India

Department of Computer Science
MTU Noida
CET-IILM-AHL
Knowledge Park-2
Greater Noida
G.B.Nagar
UP 201306,India

Abstract

The RSA cryptosystem is currently used in a wide variety of products, platforms, and industries around the world. It is found in many commercial software products and is planned to be in many more. In hardware, the RSA algorithm can be found in secure telephones, on ethernet network cards, and on smart cards.It offers encryption and digital signatures (authentication). In this paper we will illustrate the application and problem associated with RSA Algorithm.

Keywords:*RSA,Digital Signature,Cryptosystem,Public Key ,Private Key,Co-prime ,Prime Number*

1. INTRODUCTION

The RSA algorithm (1977) is widely used for public-key encryption.Developed by Ron Rivest, Adi Shamir, and Len Adleman (MIT).

The RSA digital signature has been adopted by Visa and Master Cards in the Secure Electronic Transactions (SET) standard for providing security of electronic transfers of credit and payment information over the Internet. In SET, signatures are used to provide certificates for public keys and to authenticate messages. Since public-key cryptography requires intensive computations, it is desirable to speed up these public-key computations by using either special-purpose hardware or efficient software algorithms.

2.RSA ALGORITHM:

Following steps are given below ,that are involve in RSA Algorithm.

2.1Key Generation

- Choose two distinct prime numbers, such as
- Compute $n = pq$ giving
- Compute the totient of the product as $\varphi(n) = (p - 1)(q - 1)$ giving
- Selects number e, such that $0 < e < \varphi(n)$ and e is relatively prime to $\varphi(n)$
- Compute *d*, where d = e^−1 mod φ(n).
- **Public key** is (*n*, *e*).
- **Private key** is (*n*, *d*).

2.2 RSA signing

S = m^d mod n ,Where **S** is the signature on m, m is the massage to be signed.

2.3 RSA verification

To verify that s is really the signer's signature on m, we verify if m = **S^e mod n = YES or NO**
If the result is **YES** then **S** is the signer's signature on m.

2.4 Example of RSA algorithm

We can illustrate RSA algorithm using sender (Viru) and Receiver (Puru) .Viru want to send a secure message to Puru , he perform the following steps according to the RSA Algorithm.

2.1 Key Generation

- First Viru choose two large prime numbers p and q.

Note:*Prime Number :* *A number that is not divisible by any other number than itself and 1.*

$p = 61$ and $q = 53$.

- He Compute $n = pq$ giving $n = 61 \cdot 53 = 3233$.
- Then Compute(the totient) $\varphi(n) = (p - 1)(q - 1)$ giving $\varphi(3233) = (61 - 1)(53 - 1) = 3120$.

- Now he Choose any number e where $1 < e < 3120$ that is coprime to 3120. Choosing a prime number for e leaves us only to check that e is not a divisor of 3120. Let $e = 17$.

Note:*Co- Prime:* *Two numbers are said to be relatively prime or coprime if the only number that they are both divisible by is 1.*

- He Compute d $d = 2753$.

- The **public key** is ($n = 3233$, $e = 17$). For a padded plaintext message m, the encryption function is $m^{17} \pmod{3233}$.
- The **private key** is ($n = 3233$, $d = 2753$). For an encrypted ciphertext c, the decryption function is $c^{2753} \pmod{3233}$.

2.2 RSA Signing.

Now Viru sign the message using computed Public Key =17
For instance, in order to encrypt $m = 65$,
we calculate $S = 65^{17} \pmod{3233} = 2790$.

2.3 RSA Verification.

Now Puru receive the encrypted message and he verify the signature by decrypting the signed message using computed Private Key=275

Where **S**= 2790, we calculate

$m = 2790^{2753} \pmod{3233} = 65$.

Hence the value of original message (m) is=Value of decrypted message that is means **YES** Thus **S** is the signer's signature on m.

3. PROBLEM ASSOCIATED WITH RSA ALGORITHM

The RSA algorithm suffers from the following weaknesses:

3.1 Multiplicative Property

The RSA signature scheme has the following multiplicative property, sometimes referred to as the *homomorphic* property.

If $S_1 = m_1^d \bmod n$

and $S_2 = m_2^d \bmod n$

are the signatures on messages m1 and m2 then

$$S_1 S_2 \bmod n = (m_1 m_2)^d \bmod n$$

On getting two different signed messages from a person it would be computationally feasible to derive the person's private key (d). This is because in this case, the values of S_1, S_2, n, m_1 and m_2 are known.

3.2 Integer Factorization

If an adversary is able to factor the public modulus n of someone then the adversary can compute $\varphi(n)$(Totient) and then, using the extended Euclidean algorithm, deduce the

private key d from $\varphi(n)$ and the public exponent e by solving

$ed = 1 \bmod \varphi(n)$

This constitutes a total break of the system. To guard against this p and q must be sufficiently large numbers so that factoring n is a computationally infeasible task.
However, with the rapid enhancement in computational power of modern computers it would be difficult to guarantee the computational infeasibility of factorization of large numbers.

4. CONCLUSION

This will be disastrous for the entire public key infrastructure sought to be implemented in India with the licensing of the Certifying Authorities. A breakdown of the RSA algorithm would mean that forging of the digital signatures of a Certifying Authority would be computationally feasible. This would result in the generation of fake digital signature certificates, thus defeating the very purpose of the appointment of certifying authorities and hence a public rejection of E-commerce.

Secondly, if due to a technological or mathematical breakthrough, factorization of large numbers becomes computationally feasible, the strength of the asymmetric crypto system would be shattered.

This can be achieved by removing all references to asymmetric crypto system, hash function, public key, private key etc from the legislation. Moreover the term digital signature should be replaced by the term electronic signature and this term must have a very wide definition.

REFERENCES

[1]. Niels Ferguson and Bruce Schneier, Practical Cryptography, Wiley, 2003. IEEE P1363 Standard Specifications for Public Key Cryptography, IEEE, November.
[2]. Alfred Menezes, Paul C. Van Oorschot, Scott A. Vanstone. (October 1996), Handbook of Applied Cryptography, CRC Press.
[3]. R. Rivest, A. Shamir and L. Adleman, A Method for Obtaining Digital Signatures and Public-Key Cryptosystem. Communications of the ACM, 21 (2), pp. 120-126, February 1978.
[4]. Clifford Cocks, A Note on 'Non-Secret Encryption', CESG Research Report, 20 November 1973. 1993.
[5]. RSA Laboratories, PKCS #1 v2.1: RSA Encryption Standard, June 2002.
[6]. William Stalling, Book 'Cryptography and Network Security'
[7]. D. E. Denning. Cryptography and Data Security. Addison-Wesley, Reading, MA, 1982.
[8]. T. ElGamal. A public-key cryptosystem and a signature scheme based on discrete logarithms. In *IEEE Trans. Inform. Theory*, Vol. IT-31, pp. 469-472, July, 1985
[9]. R. Gennaro, D. Katz, H. Krawczyk, T. Rabin: Secure network coding over the integers. PKC 2010.
[10]"Applied Cryptography", Second Edition, Schneider, 1996.

AUTHORS PROFILE

1. **First Author**:**Mr.Virendra Kumar**

(Assistant Professor in CS Department),Educational Qualification :B.Tech(Computer science and Engg.) from UPTU,Lucknow,M.Tech(Computer science and Engg) from Jamia Hamdard University New Delhi), worked as Network Administrator in NIIT Lucknow,Current Employer is CET-IILM-academy of Higher Learning,Gr. Noida India,My achievements are Certification in VPN from HCL Infinate Ltd. Lucknow India, academic achievements: two books are Publised on title '*Computer Organization' &' Information Security and Cyber Law'*.Currently working on GIS based Cloud Computing.

2. **Second Author**:**Mr. Puran Krishna Kaul**

(Assistant Professor in CS Department),Educational Qualification :M.Sc(Information technology.) Sikkim M.Tech(Spl Language Technology) from Centre For Devlopment Of Advance Computing Noida NCR), worked as Software developer at CDAC for nearly 2 years Noida current employer is CET-IILM-academy of Higher Learning,Gr. Noida India,acedamic achievements: paper presented "verstality of indian information and technology act 2000".

Social Networks Research Aspects : A Vast and Fast Survey Focused on the Issue of Privacy in Social Network Sites

Mohammad Soryani
Mazandaran University of Science and Technology
Mazandaran, Iran

Behrooz Minaei
Iran University of Science & Technology
Tehran, Iran

Abstract— **The increasing participation of people in online activities in recent years like content publishing, and having different kinds of relationships and interactions, along with the emergence of online social networks and people's extensive tendency toward them, have resulted in generation and availability of a huge amount of valuable information that has never been available before, and have introduced some new, attractive, varied, and useful research areas to researchers. In this paper we try to review some of the accomplished research on information of SNSs (Social Network Sites), and introduce some of the attractive applications that analyzing this information has. This will lead to the introduction of some new research areas to researchers. By reviewing the research in this area we will present a categorization of research topics about online social networks. This categorization includes seventeen research subtopics or subareas that will be introduced along with some of the accomplished research in these subareas. According to the consequences (slight, significant, and sometimes catastrophic) that revelation of personal and private information has, a research area that researchers have vastly investigated is privacy in online social networks. After an overview on different research subareas of SNSs, we will get more focused on the subarea of privacy protection in social networks, and introduce different aspects of it along with a categorization of these aspects.**

Keywords- Social Networks; Privacy; Privacy in Social Networks; SNS; Survey; Taxonomy;

I. INTRODUCTION

In recent years several attractive and user-friendly facilities have been introduced to online society and we see an extensive and increasing participation of people in various online activities like several kinds of content publishing (blogging, writing reviews etc.) and having different kinds of relationships and interactions. The huge amount of information that is generated in this way by people has never been available before and is highly valuable from different points of views. An outstanding phenomenon that has had a significant influence on this extensive participation and includes a large part of generated information is SNSs (Social Network Sites). Maybe in past, to study about the relationships, behaviors, interactions, and properties of specific groups of people it was necessary to make a lot of effort to gain some not very detailed information about them, but in the new situation and with the emergence of online social networks, and the huge amount of various activities that are logged by their users, the desired information is accessed much more simple and with incomparably more details than before by researchers. This has led to different kinds of research with different goals which we will have an overview on in this paper. The benefits and stakeholders that may benefit from having this information or having the results of analyzing it are several but some of them are: commercial companies for advertising and promoting their products, sociologists to analyze the behavior and features of different societies, intelligence organizations to prevent and detect criminal activities, educational and cultural activists for promoting their goals, employers for acquiring information about job seekers, and generally any kind of information with any application that you may think of, related to people and human societies, may be obtained by having access to the information available on SNSs or the results of analyzing these information. In this paper we try to review some of the accomplished research on the available information of SNSs and present a categorization of research topics and subareas related to online social networks' information.

As a lot of peoples' published information is private and on the other side as we will see, having access to them has a lot of applications and benefits for different parties, letting them to be available with unlimited access has consequences that sometimes may be catastrophic. We will pay more attention to this issue in this paper.

In the following sections of this paper we will first introduce sixteen research subareas about online social networks while mentioning a few of accomplished studies related with them. After that we will have a more focused review on another important research subarea namely privacy protection in social networks, and will present a categorization of its different aspects. We will conclude at the end. To have a look at the whole picture of the categorization from above, Fig.1 shows several research sub-topics about SNSs and Fig.2 extends the topic of privacy and presents a categorization of several aspects of privacy in social networks.

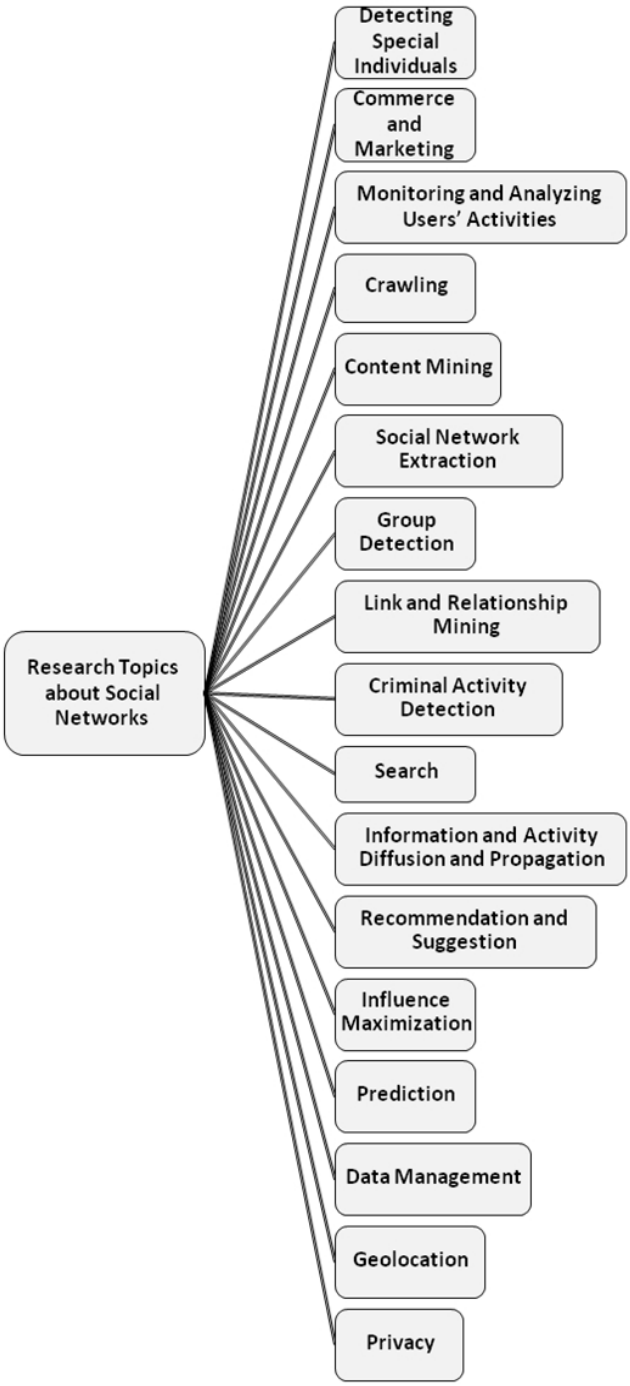


Figure 1. A categorization of research topics in social networks area.

II. A CATEGORIZATION OF RESEARCH TOPICS IN SOCIAL NETWORKS AREA

1) Detecting Special Individuals: Some people with special characteristics may be attractive for some companies, manufacturers, organizations, etc. for example it may be desirable to find special persons with high skill in a special field or to find most influential persons in propagating some special kind of content. In [3] a social search engine named Aardvark (http://vark.com) is presented that needs to find the best person for answering a specific question, and one of its information resources is people's profiles on facebook. In [4] some work is done towards forming a team of experts from members of a social network. In [5] to specify influential persons within Twitter, ranking people based on their followers, PageRank and number of retweets is investigated. In [6] some definitions are defined for different people whose actions impacts on making the same actions by others and such people are called leaders. In this paper some algorithms are presented for detecting these people by the use of a social graph and a table which contains users' actions. In [14] some references are cited in which some methods for extracting most important (central) members are presented. It has mentioned strengths and weaknesses of some metrics. In [21] Content Power Users (CPUs) in blog networks are defined as users whose published content has a lot of impact on other users' actions. In this paper a method for identifying these users is presented and some other research works about detecting highly influential people in social networks are cited.

2) Commerce and Marketing: Advertisement in SNSs can be targeted [11] [31] [6]. Targeting users whose activity influences others could be beneficial for companies [6]. As is mentioned in [11] a manufacturer can select a number of users and give them its product with some discount or even free, and hope that their influence on other users promotes their product. In the case of discounting the amount of discount is exposed to discussion. In [31] using users' profile information and the information about their activities towards targeted advertisement is mentioned.

3) Monitoring and Analyzing Users' Activities: In [12] users' behavior in a social network is analyzed to identify the patterns of closeness between colleagues. Paper [24] notes the importance of awareness about users' participation patterns in knowledge-sharing social networks for researchers and social network industry; and analyses users' activities in three social networks. Some results that are different from common assumptions are reported.

4) Crawling: To analyze the information of SNSs first we must acquire it. One of the most important ways to do this is using crawlers. Crawlers generally should have some specifications like being up to date and having mechanisms to prevent fetching the same page more than once. According to special characteristics of social networks like the huge size, and the auto crawling prevention mechanisms that SNSs use it seems that we need special kind of

crawlers. In [10] noting the large size of data, and the different way of data presentation in social networks a parallel crawler is proposed for crawling social networks. In [39] [7] facebook is crawled and the problem of facing with CAPTCHAs is noted. In [38] the ethicality of web crawlers is discussed.

5) Content Mining: One kind of information that is made by users is the content that they put in the sites in different ways. In [8] according to the real-time nature of twitter, an algorithm for monitoring and analyzing the tweets is proposed towards detecting a specific event. In this work a system is implemented to detect earthquakes in Japan and is able to do so with high accuracy. In [9] new type of texts that are published on SNSs and are usually short with an informal form of writing (called social snippets) are investigated and some applications of analyzing such texts are mentioned. Their focus is on keyword extraction from this kind of texts. In [36] mentioning the applications of identifying the quality of users' reviews about different issues, using social networks information to improve reviews quality identification is discussed. In [22] the necessity of applying automated language analysis techniques towards security in digital communities including social networks is mentioned and an approach is proposed for detecting a special kind of malicious activity.

6) Social Network Extraction: There are a lot of data in various forms on the web that apparently do not have the structure of a social network but with some mining activities on them like extracting the identity of data owners and the relations between them it is possible to extract the social network that relies beneath these data. Examples of such studies are [1] and [2]. In [1] the information of a message board is used and in [2] a system named ArnetMiner [http://www.arnetminer.org] is presented that extracts a social network of researchers. In [21] extraction of social network using a blog is mentioned.

7) Group Detection: In [4] identifying a group of skillful people to accomplish a specific job with a minimum communication cost in a social network is discussed. In [25] an efficient algorithm for large social networks named ComTector is presented for detecting communities. In [33] grouping in a social network is done towards detecting the backbone of a social network.

8) Link and Relationship Mining: In [12] relationship closeness is investigated based on the behavior of users in a social network inside a company. It is mentioned that Some behaviors are a sign of professional closeness and some are a sign of personal closeness. In [13] an approach is presented for estimation of relationship strength, and is evaluated with facebook and linkedin data. In [34] an approach for inferring the links that exist but are not observed is presented and a good survey about link mining in social networks is cited.

9) Criminal Activity Detection: Some specifications of SNSs like presence of great number of various people and new ways of communication has attracted criminals to use them for their malicious activities, so to prevent and detect such activities some special research is necessary to be done. In [22] some challenges about law enforcement and the necessity of using automated language analysis techniques for active policing in digital communities is mentioned. It notes some applications of using these techniques like identifying the child predators who pretend to be a child. In [29] a system is proposed for identifying suspects with the help of social network analysis (SNA). In [16] the application of SNA in criminal investigation and yet protecting privacy at the same time is discussed. In [32] the importance of clustering web opinions from intelligence and security informatics point of view along with some criminal activities that could be done in this space are mentioned and a clustering algorithm for detecting the context of the discussions available at social networks is presented.

10) Search: Search engines, both general purpose and special purpose ones use different information as parameters for ranking their search results. SNSs may be a valuable source of information to improve the ranking. For example special characteristics of people acquired from social networks, may be considered for ranking search results tailored to each individual's characteristics resulting in personalized search. In [19] using the structure of a social network toward improved result ranking in profile search is studied and using the social graph for improving ranking in other types of online search is mentioned as a future research. In [3] a search engine is introduced that instead of looking for appropriate documents related with the given query, it looks for suitable persons for answering the given question. To do that it gathers information about people from different resources including SNSs. In [37] towards leveraging the information about searchers in social

networks for document ranking, a framework named SNDocRank is presented. Also a study about personalizing search results using users' information is cited.

11) Information and Activity Diffusion and Propagation: The way that information and activities propagate through a social network is another area that is worth to investigate and studying about it can have various benefits. For example commercial companies may be interested in the results of such studies to improve the spread of information about their products, or educational and cultural activists may benefit from it for promoting their goals. In [20] information propagation in blogs is studied and some applications for such a study are mentioned. In [6] the spread and prevalence of users' actions is investigated over time, to identify users whose actions have influence on other users' actions.

12) Recommendation and Suggestion: According to [27] the goal of a recommender system is to recommend a set of items to a user whose favorite items are similar to them. In this work some algorithms have been designed and implemented for such systems in social networks. In [28] some techniques are presented for link prediction and the application of such techniques in friend recommendation in social networks is mentioned. In [23] some important aspects of research related with social recommendation that could be done are mentioned.

13) Influence Maximization: The problem here is to find a number of persons whose scope of influence is maximum; for example a company that has developed an application for a social network and wants to market it on that social network and can afford to invest on limited number of users (for example for giving gifts to them) would like to choose these users so that the extent of final influenced users is maximum [35]. In [26] assigning roles to users and application of being aware of these roles in influence maximization is mentioned. In [35] in addition to improving another algorithm called greedy, some heuristics are presented that run much faster.

14) Prediction: By studying the information of SNSs it is possible to predict some events that may happen in future. For example some research has been done recently to predict future links in social networks [34]. In [28] some techniques for predicting the links that may be established in future are presented.

15) Data Management: Managing the huge volume of SNSs' data has several aspects and due to special features of this kind of data, needs specific research. For example the structure of storing data is very important and can affect the amount of needed storage. In [30] a study is done about compressing social networks and a new method for compressing social networks is presented. Some of the similarities and differences between this problem and the problem of compressing web graph are found. Their results show vast difference between compressibility characteristics of social networks and web graph.

16) Geolocation: Detecting the location of the user can have several applications including personalization. Among the studies in which SNSs are used to detect the location of users is [18]. In this work researchers of facebook have mentioned some applications of knowing the location of users like news personalization, and with pointing to inconsistency of results when using ip address for geolocation they have presented a method for detecting the location of the user using information about the location of her friends. It is also mentioned that their algorithms could be run iteratively towards identifying the location of most users who have not provided any information about their location.

17) Privacy: As the focus of this paper is on the area of privacy in SNSs, we will present a broader overview on several aspects of this area along with introducing a taxonomy of these aspects which is shown in Fig.2.

17-1 Defence

17-1-1 Helping Users

Users need to be informed about the consequences of their various information publication and their activities; need to know which part of their information is accessible and for whom; need to have some facilities to control the way that people can access their information, and need to get all of these requirements in a simple and understandable way that does not need a lot of time and effort. A user gets acquainted with the matter of privacy from the first steps of her experience on SNSs by facing with privacy policies text. Definitely you too have faced with such a text and confirm that they are not very pleasant for users and a lot of people accept and pass over them without reading. In [42] the challenge of showing users' privacy related issues to them in an understandable way is presented with an interesting example where a possible interpretation of a privacy-related text may be different from what really happens. In a study that is done on six well-known SNSs [51], it is mentioned that privacy policies often have internal inconsistencies and also there is a lack of clear phrases about data retention.

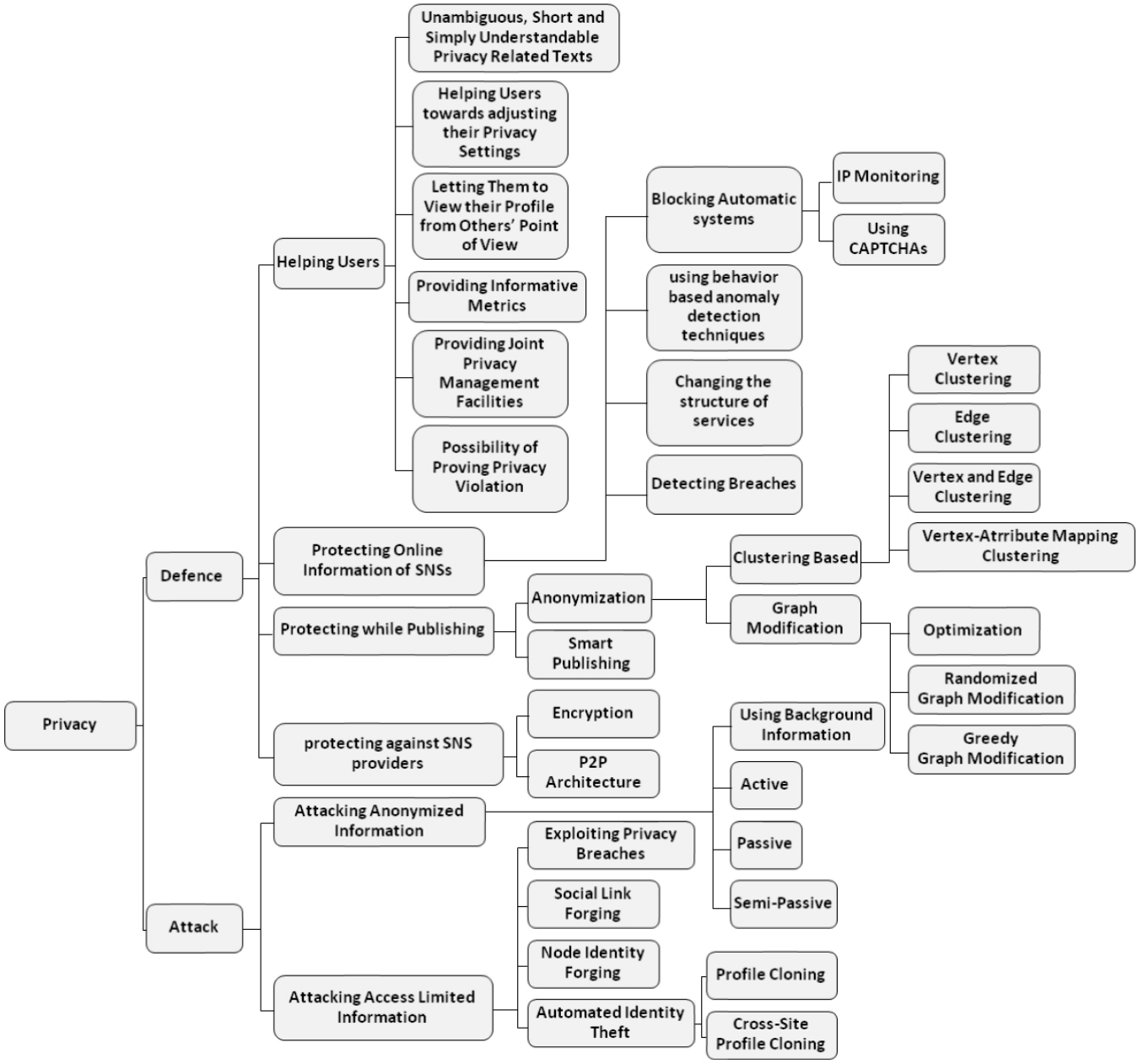


Figure 2. A categorization of privacy related aspects in Social Network Sites.

Overall privacy related texts like privacy policies should be unambiguous, as short as possible, and simply understandable. This can be considered as the first step of helping users towards their privacy protection. After accepting the privacy policy agreement, there are some flexible privacy settings that could be adjusted and help the user to specify the way her information is accessible. Since it is difficult for the average user to adjust these fine grained and detailed settings [44], some research towards helping users to do that in the best way could be useful. In [44] a work has been done toward automatically adjusting these settings with the minimum effort of the user. In this study a wizard is introduced that builds a classifier based on the user's answers to the systems requests and uses it to automatically adjust the settings. The possibility of using a limited amount of user's input to build a machine learning model that concisely describes the privacy preferences of a user, and using this model for automatically adjusting privacy settings of a user is presented in this paper. Some other interesting points about this approach are the adoption of the system after a new friend is added by a user (it presents some information about the user's preferences), and the possibility of viewing and modifying the obtained model by advanced users. Another helpful facility (the work

is presented by Lipford et al. and is cited in [44]) is a tool that makes the user able to view her own profile from the view point of each of her friends and consequently observing the results of her settings. It seems that some sort of this approach is adopted by facebook [44]. In [54] a method for suggesting privacy settings to newcomers is presented and the importance of these primitive settings due to the users' tendency towards keeping them is noted. In this study to some extent, a review about how to use machine learning to prepare and suggest primitive privacy settings that are more probably useful for users is presented. Another helpful aid for users' better privacy protection is providing some informative metrics by which users can obtain concise useful information. In [42] it is suggested to use an approach which uses data mining and potentially other AI tools to find out the amount of difficulty that accessing users' information has, and to provide some metrics for showing that, that is a useful tool for informing users about their privacy risk. Another metric that is presented in [65] is "privacy score". In this study some models and algorithms have been presented for calculating this metric and some mathematical models are developed to estimate the sensitivity and visibility of the information, which are influential on privacy risk. Sometimes users do not behave appropriately regarding their privacy while using SNSs. Understanding these risky behaviors, their causes, and then informing people and developing preventive mechanisms could be another way of helping users. In [47] a few of these risky behaviors are mentioned : incaution in accepting friendship requests, clicking on links received from others without enough caution, reacting to suspicious friendship requests after accepting them (not before) and therefore letting suspicious users to access information, interaction with fake profiles and overall, high implicit trust that exists in SNSs. A study is cited there in which 41% of 200 users whom a friendship request were randomly sent to, accepted it, and most of the users had not limited access to their personal profile information. In another study which is also cited in [47], publicly available information of some people has been gathered from some SNSs, and have been used in phishing emails; results show that targeted people whose received emails contains some information about them or their friends are more likely to get involved. In [46] [50] and [59] some surveys have been conducted as a way for studying users and acquiring information that leads to better helping them. In [46] interviews and surveys are used to investigate the effective factors that cause personal information exposure by students on facebook and why they do it despite the existing concerns. How to defend them against privacy threats is also studied. In [59] a survey is conducted in which the participants are highly educated and some information about users' behaviors, viewpoints, and concerns about privacy related issues are gained. Notwithstanding all of the existing threats and risks about private information disclosure, however users extensively tend to engage in social relations and interactions in SNSs and naturally each of these relations and interactions needs some information exposure. In [40] a study is done with the goal of determining the least information that needs to be shared to accomplish a specific interaction. An interesting study towards preparing helpful tools for users to protect their privacy is [53] in which helping users to jointly manage the privacy of their shared content is studied. Sometimes people who benefit from or get harmed by publication of a specific information unit are multiple, and publishing such information may put the privacy of several people at risk. An example of this kind of information that is extensively being published over SNSs is a photo in which several people exist. Photos are sometimes tagged and/or some additional information about them is available besides them. Support for common ownership in SNSs, and the requirements of solutions namely being fair, lightweight, and practical are mentioned as some issues that exist about joint management of privacy in [53]. In [59] it's been tried to find a way for a user to express her privacy preferences, and a method to do this with the aim of being easy to understand by other users and also being machine readable (so that it can be used by other service providers and third parties) is presented. It is mentioned that they intend to use cryptographic techniques to provide the possibility of proving privacy violation for the data owner (for example to a legal authority) which is another helpful facility towards users' privacy protection.

17-1-2 Protecting Online Information of SNSs

Let's assume that we have given all of the helpful information to users to properly manage their information publishing and protecting their privacy. Besides that we have provided them with best tools for adjusting their privacy settings in a fast, accurate, and simple way. Is it the time to relax and feel like we have done all we could do, or there are other issues that we should take care of and study about? The problem is that some of those who are interested in having users' information do not give up and try to exploit from any possible way (sometimes legal and sometimes illegal) to acquire their desired information. In addition to studies that have been done towards directly helping users, it is necessary to make some efforts to keep the online information out of undesired reach. In this part we will take a look at this issue. A study on five Russian SNSs [50] shows that despite more concerned users, there is a significant gap between their providers and western SNS Providers about understanding the privacy related concepts and preparing appropriate defencive mechanisms; and overall the privacy condition is reported to be catastrophic that leads to exposure of a large amount of users' information. An important tool that is used to acquire the online information of SNSs is a bot, with which intruders are able to accomplish their actions automatically and with a large scale. Among the actions that could be done with these automatic systems are: crawling, creating fake profiles, and sending forged friendship requests. Obviously something

needs to be done to stop these systems. In fact SNS providers should somehow determine whether the requests are sent by a real person or by an automatic system. One thing that could be done is to monitor IP addresses. Another method that is common for online automatic activity detection is using CAPTCHAs [http://www.captcha.net, 47]. A CAPTCHA is a program that protects websites against bots by creating tests that human can pass but programs cannot [http://www.captcha.net]. According to [47], facebook uses an adoption of reCAPTCHA (a related reference is cited) solution which is developed at Carnegie Mellon university. Automatic systems may have mechanisms to pass CAPTCHAs. Also they can exploit the possibility of changing CAPTCHAs that some websites offer, to find a CAPTCHA which they are able to pass. To deal with this, the rate of presented CAPTCHAs could be limited. [47]

In [50] it is concluded that most Russian SNSs do not prevent automatic profile crawling appropriately. Also according to [47] in some cases there is a lot of improvement possibility to make CAPTCHAs more difficult to break, and not every SNSs try enough to make automatic crawling and access more difficult. Another way of prevention and detection of suspicious activities is using behavior based anomaly detection techniques that can reduce the speed of the attack and its economical feasibility [47]. Although SNSs should try to protect the privacy of users and keep their information accessible only in the way that they have determined by their settings, however there may be some privacy breaches. In [47] an interesting example is mentioned. It is said that according to similar characteristics of SNSs, an extendable model could be built and potential breaches could be detected by formal analysis of this model. Changing the structure of services may help to increase the privacy protection level for users. For example in [67] extending link types is mentioned as a helpful action for privacy protection. It means instead of simply just being connected or not connected, people could for example specify the direction of their connection or the amount of trust which exists in the relationship. An example of its benefit is when a malicious user succeeds to fool another user and establish a connection with her. In this situation using trust degree can decrease the consequences of this connection. [67]

A kind of extended link types is recently adopted by google's SNS [http://plus.google.com].

In [59] and [60] using cryptographic techniques is proposed to prove privacy violations, that can prevent unauthorized access, and if happened compensate or decrease the damage.

17-1-3 Protecting while Publishing

17-1-3-1 Anonymization

So far we have discussed privacy protection by means of limiting access to users' personal information. In this section we are going to take a look at a kind of protection that aims to protect the privacy and publish information at the same time. As we mentioned, there is a high interest and desire with various motivations to have access to SNSs data. A method that is used to publish this data and protect users' privacy at the same time is called anonymization. An informal definition of anonymization in the context of privacy protection is replacing information that its revelation may damage users' privacy (email, address etc.) with other harmless data. In [36] and [52] the tradeoff between privacy and utility of anonymized data is discussed. A good survey about anonymizing social network information is [17]. In this study anonymization methods have been categorized and according to it, the state-of-the-art methods for social network information anonymization are clustering based approaches and graph modification approaches. Clustering based approaches include four subcategories of vertex clustering, edge clustering, vertex and edge clustering, and vertex-attribute mapping clustering. Graph modification approaches include three subcategories of optimization, randomized graph modification, and greedy graph modification. In [43] a weakness of past studies about anonymization is mentioned, which is their focus on methods that consider a single instance of a network, while SNSs evolutionally change and the information that does not reflect these changes is not enough for every analysis. In this paper some studies that has been done about the evolution of social networks are mentioned and also it is noted that anonymization of different instances taken in different times is not sufficient and comparing them leads to information revelation. Researchers of this paper have cited a report of their own in which they have proposed an approach for this problem. As we will see in the attacks section the beneficiaries still try to acquire their desired information and try to extract it even from anonymized data, so some studies towards improving anonymization techniques and overcoming their weaknesses like [69] have been and will be done. Researchers in [48] believe that in the area of users' privacy, mathematical rigor is needed towards having clear definitions about the meaning of comprising privacy and the information that the adversary has access to.

17-1-3-2 Smart Publishing

An application of Social Network Analysis (SNA) is criminal investigation [16] but it seems to somehow have contradiction with the matter of users' privacy. An interesting approach is presented in [16] with which without direct access to the SNS information and even their anonymized form, only some results of SNA (in form of two centrality metrics) is given to investigators and gives them the opportunity to send queries without privacy violation.

17-1-4 Protecting Against SNS Providers

Another privacy related concern is about inappropriate or undesired use of users' information by SNS providers

[67] [59] [63]. Towards solving this problem a key management scheme is presented in [15]. In the proposed method, information is encrypted and SNS providers are unaware of the keys. It is claimed that it does not have a weakness of some other related works and does not require the user and her information viewer to be present at the same time. In [55] a client-server based arichitecture is proposed for protecting users' information from SNS providers' access in which information is transferred as encrypted blocks. In [67] peer to peer (P2P) architecture for SNSs and researchers tendency for designing this architecture for next generation of SNSs is mentioned. Some of the advantages of client-server architecture over P2P architecture (like more efficient data mining in a central repository), and shortcomings of using client-server architecture with encryption (like some relations of users being detectable by other data like IP addresses) are also noted. A combination of P2P architecture and a good encryption scheme is noted as a better solution for privacy protection in SNSs.

17-2 Attack

As it was mentioned there are interested applicants that do not give up after we put some barriers on their way and try to make unauthorized information out of their reach. They still try to attack and pass the barriers and reach their desired information. In this part we take a look at these attacks from two points of view, attacks on anonymized data and attacks on access limited information.

17-2-1 Attacking anonymized Information

Despite the efforts towards protecting users' private information when publishing SNSs information using anonymization, this information is still exposed to a special kind of attacks that aim at discovering information in anonymized information. Adversaries may use some background information to accomplish this kind of attacks. [17] [52] [48]

For example the attacker may have some specific information about her target (the person who the attacker intends to get access to her information), and be able to recognize her target among the anonymized information [17]. In [48] three types of attacks are presented as active attacks, passive attacks, and semi-passive attacks : Passive attacks are described as those in which attackers begin their work to detect the identity of nodes after anonymized information is published; at the other side in active attacks the adversary tries to create some accounts in the SNS and establish some links in the network so that these links will be present in the anonymized version of the information; in semi-passive attacks there is no new account creation but some links are established with the target user before the information is anonymized. Having background information (both the information that the attackers themselves has injected to the network and the information that they have acquired in other ways) is an important tool in the hand of anonymized information attackers. In [17] some examples of background information are mentioned like attributes of vertices, vertex degree, and neighborhoods. In [69] using neighborhood information is presented as an example of a type of attacks called neighborhood attacks. In [52] vertex degree is said to be the most vastly used parameter. In [48] some studies about using interesting information like user prepared text for attacking anonymized data are cited (although in different contexts from the SN context of that paper). In [39] good information is presented about anonymization and deanonymization. In this paper a passive method is proposed for identity detection in anonymized information, and using it along with the information of two well known SNSs (twitter, flickr) notable results are obtained regarding identity detection of some members of these SNSs in the anonymized graph of twitter. In this work the information inside flickr is used as background information. An important point mentioned in [43] is that in the area of social networks anonymization, the main focus so far has been on a single instance of the network's information in a specific time, and this is inconsistent with the very dynamic nature of these networks. It is noted there that having several copies of anonymized data of a social network, which are taken in different times may lead to information revelation by comparing these copies.

17-2-2 Attacking Access Limited Information

Using the facilities which users are given, to adjust how their information could be accessed, they can make their information not to be accessible by everyone, and specify different limitations for different parts of their information, for example a person may set her pictures to only be viewable by her close friends. Attackers certainly would try to cross these borders. In [67] "social link forging attacks" and "node identity forging attacks" are mentioned. The former means deceiving a user and convincing her to establish a link (may include impersonating one of her friends by the attacker), the latter means creating several fake identities and pretending to have several personalities in a SNS. In [42] a breach is detected in linkedIn, and using the presented method the contacts of a victim are extracted. Increasing the credibility of phishers' messages using the credibility of people who are connected with the victim is noted as a motivation for this kind of attacks. Some more complicated attacks are also cited in this paper. In [47] a type of attack called "automated identity theft" along with its two subtypes called "profile cloning" and "cross-site profile cloning" is presented. In these attacks the attacker tries to make fake profiles which appear to belong to persons who really exist and have profiles either in the target SNS or other SNSs. In this study a prototype of an attack system for performing attacks is presented which performs crawling, users' information gathering, profile creation, message sending, and tries to break CAPTCHAs. Some experiments have been done on five social networks including facebook and linkedin.

III. CONCLUSIONS

Emergence of social networks and increasing participation of people in activities in these sites along with the huge amount of various information like interactions, reviews, interests and different kinds of published contents that are logged by users have attracted researchers and other parties to have access to this information or to the results of analyzing it. This information has never been available with such a huge volume, detail, and ease and speed of access before. A few number of those who are interested in having this information or the results of analyzing it alongside their motivations are: commercial companies for advertising and promoting their products, sociologists for analyzing the behavior and features of different societies, intelligence organizations for preventing and detecting criminal activities, educational and cultural activists for promoting their goals, and employers for acquiring information about job seekers. In this paper along with introducing some of the studies in this area, a categorization of research subareas was presented and a base has been provided for researchers to briefly get acquainted with some new, attractive and useful research areas. A categorization of reviewed research subtopics is illustrated in Fig.1.

People always have some private information that do not want to be exposed to public access, and if accessed by some adversaries, may have some (sometimes catastrophic) consequences. So in this survey we focused on the issue of protecting users' privacy and presented a categorization of different aspects of this area. This categorization is illustrated in Fig.2.

REFERENCES

[1] Naohiro Matsumura, David E. Goldberg, and Xavier Llor\&\#224;. 2005. Mining directed social network from message board. In *Special interest tracks and posters of the 14th international conference on World Wide Web* (WWW '05). ACM, New York, NY, USA, 1092-1093.

[2] Jie Tang, Jing Zhang, Limin Yao, and Juanzi Li. 2008. Extraction and mining of an academic social network. In *Proceeding of the 17th international conference on World Wide Web* (WWW '08). ACM, New York, NY, USA, 1193-1194.

[3] Damon Horowitz and Sepandar D. Kamvar. 2010. The anatomy of a large-scale social search engine. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 431-440.

[4] Theodoros Lappas, Kun Liu, and Evimaria Terzi. 2009. Finding a team of experts in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '09). ACM, New York, NY, USA, 467-476.

[5] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 591-600.

[6] Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan. 2008. Discovering leaders from community actionsf. In *Proceeding of the 17th ACM conference on Information and knowledge management* (CIKM '08). ACM, New York, NY, USA, 499-508.

[7] Alice Leung, Roven Lin, and Jesse Ng, Philip Szeto. 2009. Implementation of a Focused Social Networking Crawler. doi=http://courses.ece.ubc.ca/412/term_project/reports/2009/focused_social_net_crawler.pdf

[8] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 851-860.

[9] Zhenhui Li, Ding Zhou, Yun-Fang Juan, and Jiawei Han. 2010. Keyword extraction for social snippets. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 1143-1144.

[10] Duen Horng Chau, Shashank Pandit, Samuel Wang, and Christos Faloutsos. 2007. Parallel crawling for online social networks. In *Proceedings of the 16th international conference on World Wide Web* (WWW '07). ACM, New York, NY, USA, 1283-1284.

[11] Nicole Immorlica and Vahab S. Mirrokni. 2010. Optimal marketing and pricing over social networks. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 1349-1350.

[12] Anna Wu, Joan M. DiMicco, and David R. Millen. 2010. Detecting professional versus personal closeness using an enterprise social network site. In *Proceedings of the 28th international conference on Human factors in computing systems* (CHI '10). ACM, New York, NY, USA, 1955-1964.

[13] Rongjing Xiang, Jennifer Neville, and Monica Rogati. 2010. Modeling relationship strength in online social networks. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 981-990.

[14] Katarzyna Musia\&\#322;, Przemys\&\#322;aw Kazienko, and Piotr Br\&\#243;dka. 2009. User position measures in social networks. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis* (SNA-KDD '09). ACM, New York, NY, USA, , Article 6 , 9 pages.

[15] Keith Byron Frikken and Preethi Srinivas. 2009. Key allocation schemes for private social networks. In *Proceedings of the 8th ACM workshop on Privacy in the electronic society* (WPES '09). ACM, New York, NY, USA, 11-20.

[16] Florian Kerschbaum and Andreas Schaad. 2008. Privacy-preserving social network analysis for criminal investigations. In *Proceedings of the 7th ACM workshop on Privacy in the electronic society* (WPES '08). ACM, New York, NY, USA, 9-14.

[17] Bin Zhou, Jian Pei, and WoShun Luk. 2008. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *SIGKDD Explor. Newsl.* 10, 2 (December 2008), 12-22.

[18] Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 61-70.

[19] Jonathan Haynes and Igor Perisic. 2009. Mapping search relevance to social networks. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis* (SNA-KDD '09). ACM, New York, NY, USA, , Article 2 , 7 pages.

[20] Yong-Suk Kwon, Sang-Wook Kim, Sunju Park, Seung-Hwan Lim, and Jae Bum Lee. 2009. The information diffusion model in the blog world. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis* (SNA-KDD '09). ACM, New York, NY, USA, , Article 4 , 9 pages.

[21] Seung-Hwan Lim, Sang-Wook Kim, Sunju Park, and Joon Ho Lee. 2009. Determining content power users in a blog network. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis* (SNA-KDD '09). ACM, New York, NY, USA, , Article 5 , 8 pages.

[22] Danny Hughes, Paul Rayson, James Walkerdine, Kevin Lee, Phil Greenwood, Awais Rashid, Corinne May-Chahal, and Margaret Brennan. 2008. Supporting Law Enforcement in Digital Communities

through Natural Language Analysis. In *Proceedings of the 2nd international workshop on Computational Forensics* (IWCF '08), Sargur N. Srihari and Katrin Franke (Eds.). Springer-Verlag, Berlin, Heidelberg, 122-134.

[23] Irwin King, Michael R. Lyu, and Hao Ma. 2010. Introduction to social recommendation. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 1355-1356.

[24] Lei Guo, Enhua Tan, Songqing Chen, Xiaodong Zhang, and Yihong (Eric) Zhao. 2009. Analyzing patterns of user content generation in online social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '09). ACM, New York, NY, USA, 369-378.

[25] Nan Du, Bin Wu, Xin Pei, Bai Wang, and Liutong Xu. 2007. Community detection in large-scale social networks. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (WebKDD/SNA-KDD '07). ACM, New York, NY, USA, 16-25.

[26] Jerry Scripps, Pang-Ning Tan, and Abdol-Hossein Esfahanian. 2007. Node roles and community structure in networks. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (WebKDD/SNA-KDD '07). ACM, New York, NY, USA, 26-35.

[27] Zeinab Abbassi and Vahab S. Mirrokni. 2007. A recommender system based on local random walks and spectral methods. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (WebKDD/SNA-KDD '07). ACM, New York, NY, USA, 102-108.

[28] Tomasz Tylenda, Ralitsa Angelova, and Srikanta Bedathur. 2009. Towards time-aware link prediction in evolving social networks. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis* (SNA-KDD '09). ACM, New York, NY, USA, , Article 9 , 10 pages.

[29] Li Ding, Dana Steil, Matthew Hudnall, Brandon Dixon, Randy Smith, David Brown, and Allen Parrish. 2009. PerpSearch: an integrated crime detection system. In *Proceedings of the 2009 IEEE international conference on Intelligence and security informatics* (ISI'09). IEEE Press, Piscataway, NJ, USA, 161-163.

[30] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, Michael Mitzenmacher, Alessandro Panconesi, and Prabhakar Raghavan. 2009. On compressing social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '09). ACM, New York, NY, USA, 219-228.

[31] Mitra, P. & Baid, K., 2009. Targeted advertising for online social networks. *2009 First International Conference on Networked Digital Technologies*, p.366-372.

[32] Christopher C. Yang and Tobun D. Ng. 2009. Web opinions analysis with scalable distance-based clustering. In *Proceedings of the 2009 IEEE international conference on Intelligence and security informatics* (ISI'09). IEEE Press, Piscataway, NJ, USA, 65-70.

[33] Nan Du, Bin Wu, and Bai Wang. 2007. Backbone Discovery in Social Networks. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence* (WI '07). IEEE Computer Society, Washington, DC, USA, 100-103.

[34] Heath Hohwald, Manuel Cebrian, Arturo Canales, Rub\&\#233;n Lara, and Nuria Oliver. 2009. Inferring Unobservable Inter-community Links in Large Social Networks. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04* (CSE '09), Vol. 4. IEEE Computer Society, Washington, DC, USA, 375-380.

[35] Wei Chen, Yajun Wang, and Siyu Yang. 2009. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '09). ACM, New York, NY, USA, 199-208.

[36] Yue Lu, Panayiotis Tsaparas, Alexandros Ntoulas, and Livia Polanyi. 2010. Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 691-700.

[37] Liang Gou, Hung-Hsuan Chen, Jung-Hyun Kim, Xiaolong (Luke) Zhang, and C. Lee Giles. 2010. SNDocRank: document ranking based on social networks. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 1103-1104.

[38] C. Lee Giles, Yang Sun, and Isaac G. Councill. 2010. Measuring the web crawler ethics. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 1101-1102.

[39] Leyla Bilge, Thorsten Strufe, Davide Balzarotti, and Engin Kirda. 2009. All your contacts are belong to us: automated identity theft attacks on social networks. In *Proceedings of the 18th international conference on World wide web* (WWW '09). ACM, New York, NY, USA, 551-560.

[40] Eran Toch, Norman M. Sadeh, and Jason Hong. 2010. Generating default privacy policies for online social networks. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems* (CHI EA '10). ACM, New York, NY, USA, 4243-4248.

[41] Jonathan Anderson, Claudia Diaz, Joseph Bonneau, and Frank Stajano. 2009. Privacy-enabling social networking over untrusted networks. In *Proceedings of the 2nd ACM workshop on Online social networks* (WOSN '09). ACM, New York, NY, USA, 1-6.

[42] Jessica Staddon. 2009. Finding "hidden" connections on linkedIn an argument for more pragmatic social network privacy. In *Proceedings of the 2nd ACM workshop on Security and artificial intelligence* (AISec '09). ACM, New York, NY, USA, 11-14.

[43] Smriti Bhagat, Graham Cormode, Balachander Krishnamurthy, and Divesh Srivastava. 2010. Privacy in dynamic social networks. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 1059-1060.

[44] Lujun Fang and Kristen LeFevre. 2010. Privacy wizards for social networking sites. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 351-360.

[45] Arvind Narayanan and Vitaly Shmatikov. 2009. De-anonymizing Social Networks. In *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy* (SP '09). IEEE Computer Society, Washington, DC, USA, 173-187. DOI=10.1109/SP.2009.22 http://dx.doi.org/10.1109/SP.2009.22

[46] Alyson L. Young and Anabel Quan-Haase. 2009. Information revelation and internet privacy concerns on social network sites: a case study of facebook. In *Proceedings of the fourth international conference on Communities and technologies* (C\&\#38;T '09). ACM, New York, NY, USA, 265-274.

[47] Lars Backstrom, Cynthia Dwork, and Jon Kleinberg. 2007. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the 16th international conference on World Wide Web* (WWW '07). ACM, New York, NY, USA, 181-190.

[48] Slava Kisilevich and Florian Mansmann. 2010. Analysis of privacy in online social networks of runet. In *Proceedings of the 3rd international conference on Security of information and networks* (SIN '10). ACM, New York, NY, USA, 46-55.

[49] Leanne Wu, Maryam Majedi, Kambiz Ghazinour, and Ken Barker. 2010. Analysis of social networking privacy policies. In *Proceedings of the 2010 EDBT/ICDT Workshops* (EDBT '10). ACM, New York, NY, USA, , Article 32 , 5 pages.

[50] Balachander Krishnamurthy and Craig E. Wills. 2008. Characterizing privacy in online social networks. In *Proceedings of the first workshop on Online social networks* (WOSN '08). ACM, New York, NY, USA, 37-42.

[51] Anna Cinzia Squicciarini, Mohamed Shehab, and Federica Paci. 2009. Collective privacy management in social networks. In *Proceedings of the 18th international conference on World wide web* (WWW '09). ACM, New York, NY, USA, 521-530.

[52] Xiaowei Ying, Kai Pan, Xintao Wu, and Ling Guo. 2009. Comparisons of randomization and K-degree anonymization schemes for privacy preserving social network publishing. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis* (SNA-KDD '09). ACM, New York, NY, USA, , Article 10 , 10 pages.

[53] Tiancheng Li and Ninghui Li. 2009. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '09). ACM, New York, NY, USA, 517-526.

[54] Esma Aimeur, Sebastien Gambs, and Ai Ho. 2009. UPP: User Privacy Policy for Social Networking Sites. In *Proceedings of the 2009 Fourth International Conference on Internet and Web Applications and Services* (ICIW '09). IEEE Computer Society, Washington, DC, USA, 267-272.

[55] Esma Aimeur, Sebastien Gambs, Ai Ho, "Towards a Privacy-Enhanced Social Networking Site," Availability, Reliability and Security, International Conference on, pp. 172-179, 2010 International Conference on Availability, Reliability and Security, 2010

[56] Safebook A Privacy-Preserving Online Social Network Leveraging on Real-Life Trust

[57] Kun Liu and Evimaria Terzi. 2009. A Framework for Computing the Privacy Scores of Users in Online Social Networks. In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining* (ICDM '09). IEEE Computer Society, Washington, DC, USA, 288-297.

[58] Chi Zhang, Jinyuan Sun, Xiaoyan Zhu, and Yuguang Fang. 2010. Privacy and security for online social networks: challenges and opportunities. *Netwrk. Mag. of Global Internetwkg.* 24, 4 (July 2010), 13-18.

[59] B. K. Tripathy and G. K. Panda. 2010. A New Approach to Manage Security against Neighborhood Attacks in Social Networks. In *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining* (ASONAM '10). IEEE Computer Society, Washington, DC, USA, 264-269.

Hybrid Multiobjective Evolutionary Algorithms: A Survey of the State-of-the-art

Wali Khan Mashwani

Department of Mathematics, Kohat university of Science & Technology, Kohat, 26000, Khyber Pukhtunkhwa, Pakistan

Abstract

This paper reviews some state-of-the-art hybrid multiobjective evolutionary algorithms (MOEAs) dealing with multiobjective optimization problem (MOP). The mathematical formulation of a MOP and some basic definition for tackling MOPs, including Pareto optimality, Pareto optimal set (PS), Pareto front (PF) are provided in Section 1. Section 2 presents a brief introduction to hybrid MOEAs.

Keywords: Multiobjective optimization, MOP, Hybrid MOEAs.

1. Introduction

A multiobjective optimization problem (MOP) can be stated as follow: [1]

$$\textbf{Minimize } F(x) = (f_1(x),.., f_m(x))^T \quad (1)$$
$$\textbf{subject to } x \in \Omega$$

Where Ω is the decision variable space, $x=(x_1, x_2,..,x_n)^T$ is a decision variable vector and x_i, $i=1,...,n$ are called decision variables, $F(x)$: $\Omega \rightarrow R^m$ consist of m real valued objective functions and R^m is called the objective space.

If Ω is closed and connected region in R^n and all the objectives are continuous of x, we call a problem 1 is a continuous MOP.

Very often, the objectives of the problem (1) are in conflict with one another or are incommensurable. There doesn't exist a single solution in the search space Ω that can minimize all the objectives functions simultaneously. Instead, one has to find the best trade-offs among the objectives. These trade-offs can be better defined in terms of Pareto optimality. The Pareto optimality concept was 1st introduced by eminent economists Pareto and Edgeworth [1]. A formal definition of the Pareto optimality is given as follows [2, 3, 4, 5].

Definition: Let $u = (u_1, u_2,..,u_m)^T$ and $v=(v_1, v_2,..,v_m)^T$ be any two given vectors in R^m. Then u is said to dominate v, denoted as $u \prec v$, if and only if the following two conditions are satisfied.

1. $u_i \leq v_i$ for every $i \in \{1, 2,..., m\}$
2. $u_j < v_j$ for at least one index $j \in \{1, 2,..., m\}$.

Remarks: For any two given vectors, u and v, there are two possibilities:

1. Either u dominates v or v dominates u
2. Neither u dominates v nor *does* v *dominate* u.

Definition: A solution $x \in \Omega$ is said to be a Pareto optimal to the problem (1) if there is no other solution $x \in \Omega$ such that $F(x)$ dominates $F(x^*)$. $F(x^*)$ is then called Pareto optimal (objective) vector.

Remarks: Any improvement in a Pareto optimal point in one objective must lead

[1] **The minimization problem can easily convert into maximization problem by multiplying each objective with −1 and vice versa.**

to deterioration in at least one other objective.

Definition: The set of all the Pareto optimal solutions is called Pareto set (PS): $PS = \{x \in \Omega | \nexists y \in \Omega, F(y) \prec F(x)\}$

Definition: The image of the **Pareto optimal set (PS)** in the objective space is called ***Pareto front (PF),*** $PF = \{F(x) | x \in PS\}$.

Recent years have witnessed significant development in MOEAs for dealing MOPs. In last two decades, a variety of MOEAs have been proposed. The success of most MOEAs depends on the careful balance of two conflicting goals, exploration (i.e., searching new Pareto-optimal solution) and exploitation (i.e., refining the obtained PS). To achieve these two goals, hybridization is good strategy [6]. The following section introduces hybrid algorithms.

2. Hybrid Multiobjective Evolutionary Algorithms

Hybrid MOEAS or combination of MOEAs with efficient techniques have been investigated for more than one decade [7]. Hybridization uses desirable properties of different techniques for better algorithmic improvements. Hybridization can be done in several ways, 1) to use one algorithm to generate a population and then apply another technique to improve it, 2) to use multiple operators in an evolutionary algorithm, and 3) to apply local search to improve the solutions obtained by MOEAs [8].

Multiobjective memetic algorithms (MOMAs) are a special type of hybrid MOEAs. MOMAs are population-based algorithms inspired by the Darwinian principles of natural evolution and Dawkins notion of a meme (i.e., defined as a unit of cultural evolution which is capable of local refinements). They are well-known algorithms for their fast convergence speed and for finding more accurate solutions to different search and optimization problems. In the following subsections, we present some state-of-the-art MOMAS.

1. Local Search Based Multiobjective Evolutionary Algorithms

Ishibuchi and Murata 1st proposed multi-objective genetic local search algorithm (MOGLS) for solving combinatorial multiobjective optimization problems [9, 10]. MOGLS applied a local search method after classical variation operators. In MOGLS, a scalar fitness function is used to select a pair of parent solutions to generate new solutions with crossover and mutation operator.

An improved version of MOGLS [9, 10] is proposed in [11]. It applies hill climbing local search optimizer on some best individuals in its current population. Its performance was tested on combinatorial multiobjective optimization in comparison with MOGLS [9, 10], strength Pareto evolutionary algorithm (SPEA) [12, 13], NSGA-II [14] and Hybrid NSGA-II [14].

Another version of MOGLS was proposed by Jaskiewicz in [15]. The basic idea of his MOGLS is to reformulate a MOP as simultaneous optimization of all the aggregation constructed by weight sum approach or Tchebycheff approach. At each generation, it optimizes a randomly generated aggregation objective.

Pareto memtic algorithm (PMA) is suggested in [16]. It uses unbounded" current set of solutions" and from it selects a small"temporary population (TP)" that compromises the best solutions with respect to scalarizing functions. Then TP is used to generate offspring by crossover operators. Jaskiwicz suggests that scalar functions are

very good to promote diversity than dominance ranking methods [17].

In [18], a biased neighborhood structure based local search is proposed. The algorithm assigns large probabilities to the neighbors of the current solution located in the promising region of the search space. The proposed algorithm perform very well on both multiobjective 0/1 knapsack and flowshop scheduling problems.

Memetic Pareto archive evolution strategy (M-PAES) is developed in [19]. It utilizes Pareto ranking based selection and grid-type partition of the objective space instead of scalarizing functions. This modified selection scheme is much faster than the scalarizing functions which are used in Ishibush's MOGLS [9] and Jaszkeiwicz's MOGLS [20, 21]. Furthermore, M-PAES maintains two archives, one stores global nondominated solutions and the other is used as the comparison set for the local search phase. M-PAES is tested against the local search optimizer, (1+1)-PAES [22] and SPEA [12, 13] on the multiobjective 0/1 knapsack problems. M-PAES has shown better experimental results than its competitors.

In [23], a memetic algorithm is suggested for dynamic muliobjective optimization. This algorithm has incorporated two adaptive hill climbing local search methods, greedy crossover-based hill climbing local search method and steepest mutation-based hill climbing local search method.

In [24], two fitness function schemes, the weighted sum fitness function and the NSGA-II fitness evaluation, are used probabilistically. The authors used the probability to specify how often the scalarizing function is used for parent selection. When the probability becomes very low, then the proposed algorithm is almost the same as NSGA-II.

[25] Proposed a local search method which uses quadratic approximations. The solutions produced in the evolutionary process of the multiobjective genetic algorithm (MOGA) [26, 27] are utilized to fit these quadratic approximations around the point selected for local search. The proposed algorithm has shown more accurate experimental results than pure MOGA [26, 27]. The same local search is also used in [28, 29, 30]. A novel agent-based memetic algorithm (AMA) algorithm based on multi-agent concepts is suggested in [31]. AMA used different life span learning processes (LSLPs) based on several local and directed search procedures strategies such as totally random, random restricted, search directions based. In AMA, an agent chooses a LSLP as a search operator adaptively and improves its algorithmic performance. Same ideas are used in [31, 32, 33, 34].

In [35], a novel iterative search procedure, called the hill climber with sidestep (HCS) is designed. HCS is capable of moving toward and along the local Pareto set depending on the distance of the current iterate toward this set. HCS utilizes the geometry of the directional cones and works with or without gradient information. HCS used as a typical mutation operator in SPEA2

[36] and developed a MOMA denoted by SPEA2HCS. SPEA 2HCS is more effective and efficient in dealing with continuous MOPs.

Two Local search methods, Hooke and Jeeves method [37, 38, 39] and steepest descent method [40, 41], are combined with S-Metric Selection Evolutionary Multiobjective Algorithm (SMS-EMOA) [42] and its two hybrid versions, Relay SMS-EMOA hybrid and Concurrent SMS-EMOA hybrid are developed in [43]. Steepest descent method used in Relay SMS-EMOA hybrid and Hooke and Jeeves method used in Concurrent SMS-EMOA hybrid. Experimental analysis on academic test functions [44] show increased convergence speed as well as improved accuracy of the

solution set of these new hybridizations.

A novel searching algorithm called the multiple trajectory search (MTS) is developed in [45]. The MTS uses multiple agents to search the solution space concurrently. Each agent does an iterated local search using one of four candidate local search methods. By choosing a local search method that best fits the landscape of a solution's neighborhood, an agent may find its way to a local optimum or the global optimum. MTS is tested on multiobjective optimization test problems designed for CEC'09 special session and competition on performance assessment of multiobjective optimization algorithms [46]. In [47], MTS is tested on CEC'09 test instances [48]. In [49], a novel Lamarckian learning strategy is designed and hybrid version of nondominated neighbor immune algorithm [50] called multi-objective lamarckian immune algorithm (MLIA) is proposed. The Lamarckian learning performs a greedy search which proceeds towards the goal along the direction obtained by Tchebycheff approach and generates the improved progenies or improved decision vectors, so single individual will be optimized locally and the newcomers yield an enhanced exploitation around the nondominated individuals in less-crowded regions of the current trade-off front. Simulation results based on twelve benchmark problems show that MLIA outperforms the original immune algorithm and NSGA-II in approximating Pareto-optimal front in most of the test problems. When compared with the state of the art algorithm MOEA/D, MLIA shows better performance in terms of the coverage of two sets metric, although it is laggard in the hyper volume metric.

A new hybrid line search approach called the Line search generator of Pareto frontier (LGP) is developed in [51]. The framework of the LGP consist of two phases, Convergence phase and spreading phase. It has been tested on OKA1 and OKA2 test problems [52], DTLZa and DTLZb test problems [53] and VLMOP2 and VLMP3 test problems [54].

2.2. Hybrid MOEAs Based on Pareto Dominance

In [55], two well-known Pareto dominance based algorithms, SPEA2 [36] and NSGA-II [14], combined with probabilistic local search and developed its hybrid versions for dealing combinatorial multiobjective optimization. In both hybrid algorithms, the use of the Local search is terminated when no better solution to current solution is found in its k neighbors. One potential advantage of proposed hybrid algorithms over its pure versions is the decrease in the CPU time.

T. Murata et al. generalized the replacement rules based on dominance relation for multiobjective objective optimization in [56]. Ordinary two-replacement rules based on dominance are usually employed in the local search for multiobjective optimization. One rule replaces a current solution with a solution which dominated it. The other rule replaces the solution with a solution which is not dominated by it. The movable area with 1st rule is very small when the number of objectives is large. On the other hand, it is too huge to move efficiently with second rule. The authors generalized these extreme rules by counting the number of improved objective values. Proposed local search based on generalized replacement rules is incorporated in SPEA [12, 13] and developed its hybrid SPEA.

In [57], two hybrid MOEAs, hybrid NSGA-II and hybrid SPEA2 are developed. In both proposed hybrid algorithms, a convergence acceleration operator (CAO) is used as an additional operator for improving the search capability and convergence speed. CAO is applied in the objective space for improved solutions. The improved objective vectors

are then mapped back to the decision space to predict their corresponding decision variables. In [58], three local search methods: simulating annealing (SA), tabu search (TS), and hill climbing local search method, are combined with multi-objective genetic algorithm [26, 27]. MOGA with hill climbing local search method has found much better approximated set of solutions on ZDT test problems [44] than pure MOGA [26, 27] and others hybrid versions of MOGA.

A sequential quadratic programming (SQP) coupled with NSGA-II [14] in [59, 60] for solving continuous MOPs [46]. The same idea is used in [61]. In [62], hybrid version of NSGA-II is suggested which combines a local search method with NSGA-II [14] for estimating the nadir point.

In [63], SQP as a local search method based on augmented achievement scalarizing function (ASF) [64] is used in the framework of NSGA-II [14] for solving ZDT benchmarks [44, 53]. SQP is also used as local search method with NSGA-II [14] as global search method in [65] and solved the CEC'9 test instances [48] in effective ways.

Hybrid constrained optimization evolutionary algorithm (HCOEA) is proposed for constrained optimization in [66]. HCOEA used niching genetic algorithm (NGA) based on tournament selection as a global search method and the best infeasible individual replacement scheme as local search operator. NGA effectively promotes the diversity in its population and local search model remarkably accelerates the convergence speed of the HCMOEA.

In [67], a fuzzy simplex genetic algorithm (FSGA) is developed. The proposed method uses fuzzy dominance concept and simplex-based local search method
[68] for solving continuous MOPs. The performance of the FSGA is more effective than NSGA-II [14] and SPEA2 [36] on ZDT test problems.

A Pareto Following Variation Operator (PFVO) is used in NSGA-II [14] an additional operator and designed hybrid NSGA-II in [69]. PFVO takes the available objectives values in the current nondominated front as inputs and generates approximated design variables for the next front as the output. The Proposed algorithm has obtained much better set of optimal solutions to ZDT test problems [44]. PFVO is also used in SPEA2
[36] and in regularity model-based multiobjective estimation of distribution algorithm (RM-MEDA) [70] and suggested its hybrid algorithms in [71]. Experimental analysis raveled that both hybrid algorithms PFVO has enhanced the convergence ability of SPEA2 [36] and RM-MEDA [70] on ZDT test problems [44, 53].

Recently, hybrid version of Archive-based Micro Genetic Algorithm (AMGA) [72] is developed in [73]. In this algorithm, SQP is used as a mutation operator genetic mutation. The inclusion of SQP speeds-up the search process of the proposed hybrid AMGA. Hybrid AMGA has found global Pareto-optimal front and the extreme solutions on most CEC'09 test instances [48]. In [74], the functional-specialization multi-objective real-coded genetic algorithm (FS-MOGA) is proposed. FS-MOGA adaptively switched two search strategies specialized for global and local search. This algorithm chooses an individual from the current population at random. If the chosen individual is a non-dominated solution, then it executes the local search procedure. Otherwise, it performs the global search procedure.
In [75], a hybrid NSGA-II is developed to deal with engineering shape design problems with two conflicting objectives: weight of the structure and maximum deflection of the structure. This algorithm used hill climbing local search method.
In [76, 77], hybrid strategy based on two-stage search process is developed for solving

many-objective optimization. The first stage of the search is directed by a scalarization function and the second stage by Pareto selection enhanced with adaptive Q-Ranking.
In [78, 79], a hybrid version of NSGA-II [14] called NSS-GA is proposed for solving ZDT test problems
[44] and DTLZ [53] test problems. NSS-GA used two direct search methods, Nelder and Mead's method [68] and golden section algorithm, for improving.
A new hybrid MOEA, the niched Pareto tabu search combined with genetic algorithm (NPTSGA) is presented dealing with multi-objective optimization problems [80]. The NPTSGA is developed on the thoughts of integrating genetic algorithm (GA) with the improved tabu search (TS) based MOEA, niched Pareto tabu search (NPTS). The proposed NPTSGA is then tested through a simple test example and compared with other two techniques, NPTS and niched Pareto genetic algorithm (NPGA). Computational results indicate that the proposed NPTSGA is an efficient and effective method for solving multi-objective problems.
A hybrid algorithm with on-line landscapes approximation for expensive MOPs, called, ParEGO is developed in [81, 82]. ParEGO is an extension of the single-objective efficient global optimization (EGO) [83]. It uses a design-of-experiment inspired initialization procedure and learn a Gaussian processes model of the search landscape, which is updated after every function evaluation. ParEGO generally outperformed NSGA-II [14] on the used test problems.

2.3. Enhanced Versions of MOEA/D

Recently, an efficient framework known as MOEA/D: multiobjective evolutionary algorithm based on decomposition, is developed in [84]. This generic algorithm bridges decomposition techniques and evolutionary algorithms. MOEA/D decomposes a MOP into many different single-objective sub problems (SOPs) and defines neighborhood relations among these sub problems. The objective of each sub problem is a weighted aggregation of the original objective functions. Each SOP is optimized by using information, mainly from its neighborhood sub problems. The SOPs in one neighborhood are assumed to have similar fitness landscapes and their respective optimal solutions are most probable be close to each others. This section provides some latest versions of MOEA/D [84].

In [85], 2-opt local search method is combined with MOEA/D [84] and tested on multiobjective traveling salesman problems (m-TSPs).

Two neighbourhoods are used and a new solution is allowed to replace a very small number of old solutions in [86]. The proposed algorithm denoted by MOEA/D-DE and tested on continuous test MOPs with complicated PS shapes [86]. MOEA/D-DE has shown much better algorithmic improvement than NSGA-II [14].

Recently, another important version of MOEA/D [84], called massively multi-topology sizing of analog integrated circuits is developed in [87]. In this version, each sub problem records more than one solution to maintain diversity.

In [88], an idea of simultaneously using different types of scalarizing functions in MOEA/D is proposed aimed to overcome the difficulty in choosing an appropriate scalarizing function for particular multiobjective problem. Weighted sum and the weighted Tchebycheff are used as scalarizing functions. Two implementation schemes of the proposed idea are examined in this paper. One is to use multiple grids of weight vectors where each grid is used by a single scalarizing function. The other is to use different scalarizing functions in a single grid of weight vectors where a different scalarizing function is alternately assigned to

each weight vector. The effectiveness of these implementation schemes was examined through experiments on multiobjective 0/1 knapsack problems with two, four and six objectives. Experimental results showed that the simultaneous use of the weighted sum and the weighted Tchebycheff outperformed their individual use in MOEA/D.

Another important extension of MOEA/D called MOEA/D-EGO, the Gaussian stochastic process model for expensive multiobjective optimization is proposed in [89]. At each iteration, in MOEA/D-EGO, a Gaussian stochastic process model for each subproblem is built based on data obtained from the previous search, and the expected improvements of these subproblems are optimized simultaneously by using MOEA/D for generating a set of candidate test points. Further, MOEA/D assisted by metamodel-Gaussian random field metamodel (GRFM) was proposed in [90].

Competition and adaptation of search directions are incorporated in MOEA/D and its effective hybrid version called EMOSA is developed in [91]. In EMOSA, the current solution of each sub problem is improved by simulated annealing with different temperature levels. After certain low temperature levels, to approximate various parts of the PF, a new method to tune the weight vectors of these aggregation functions is suggested. Contrary to the original MOEA/D, no crossover is performed in this hybrid approach. Instead, diversity is promoted by allowing uphill moves following the simulated annealing rationale.

In [92], MOEA/D with NBI-style Tchebycheff approach is developed. The new style Tchebycheff approach replaces the already used weighted sum approach and Tchebycheff approach. The proposed algorithm deals with disparately scaled objectives of constrained portfolio optimization problems effectively.

In [93], an enhance version of MOEA/D [84] is established. In this algorithm, 1) DE operator replaced with a guided mutation operator for reproduction, 2) a new update mechanism with a priority order is proposed. The update mechanism can improve MOEA/D's performance when the SOPs obtained by decomposition are not uniformly distributed on the Pareto font. Finally, the set of test instances for the CEC'09 competition is used for evaluating the performance of the various combinations of these mechanisms in developed approach.

A novel multiobjective particle swarm optimization based on decomposition algorithm developed in [94]. In algorithm, PSO coupled with MOEA/D [84] for solving continuous problems.

An adaptive mating selection mechanism (AMS) is introduced in MOEA/D and the resultant version is called MOEA/D-AMS. AMS consist of controlled subproblems selection scheme (CSS) and matting pool adjustment (MPA). The CSS assigns the computational efforts to different subproblems. The MPA mates individuals with those who are close on the decision space so that small change of gene values can be achieved, which are required at the late stage of evolutionary process.

A new improved version of MOEA/D [84] called, TMOEA/D is developed. TMOEA/D utilizes a monotonic increasing function to transform each individual objective function into the one so that the curve shape of the non-dominant solutions of the transformed multi-objective problem get close to the hyper-plane whose intercept of coordinate axes is equal to one in the original objective function space. In [95], two mechanisms are introduced. Firstly, a new replacement mechanism to maintain a balance between the diversity of the population and the employment of good information from neighbors; secondly, a randomized scaling factor of DE is adopted in order to enhance the search ability of MOEA/D-DE [86] on

real-world problem, the sizing of a folded-cascade amplifier with four performance objectives.

In [96], a new version of MOEA/D [84], called (MOEA/DFD) is developed. The proposed algorithm introduced a fuzzy dominance concept for comparing two solutions and used scalar decomposition method when one of the solutions fails to dominate the other in terms of a fuzzy dominance level. MOEA/DFD outperforms other MOEAs.

In [97], an interactive version of the decomposition based multiobjective evolutionary algorithm (IMOEA/D) is proposed for interaction between the decision maker (DM) and the algorithm. During the stage of interaction, IMOEA/D presents preferred sub problems to DM to choose their most favorite one, and then guided the search to the neighborhood of selected sub problems. IMOEA/D used the utility function which is modeled in [98]. The used utility function simulates the responses of the DM in IMOEA/D implementation. IMOEA/D has been handled the preference informations very well and successfully converged to the expected preferred regions.

Very recently, the behavior of MOEA/D is examined on multiobjective problems with highly correlated objectives in [99]. The performance of MOEA/D is severely degraded while SPEA 2 [36] and NSGA-II [14] had offered good behaviors on highly correlated objectives.

In [100], a novel method called Pareto-adaptive weight vectors (*paλ*) is proposed. This method automatically adjusts the weight vectors in MOEA/D [86] which are associated with each subproblem. The algorithm, called, multiobjective optimization by decomposition with (*paλ*) is tested on continuous test problems [44, 53] in comparison with simple MOEA/D [84] and NSGA-II [14] on each test problem.

The paper in [101], studies the effects of the use of two crossover operators in multiobjective evolutionary algorithm based on decomposition with dynamical resource allocation (MOEA/D-DRA) [102] for multi-objective optimization. The two crossover operators used are, simplex crossover operator (SPX) and center of mass crossover operator (CMX). The use probability of each operator is updated dynamically based on its corresponding successful reward. The experimental results showed that the use of two crossover operators in MOEA/D-DRA [102] can improve its performance on most of the CEC'09 test instances [48].

A combination of MOEA/D and NSGA-II for dealing with multiobjective CARP (MO-CARP) is proposed in [103]. The MO-CARP is a challenging combinatorial optimization problem with many real-world applications, e.g., salting route optimization and fleet management. The proposed memetic algorithm (MA) called decomposition-based MA with extended neighborhood search (D-MAENS) has shown better performances than NSGA-II [84] and and the state-of-the-art multiobjective algorithm for MO-CARP (LMOGA) [104].

In [105], a hybrid evolutionary metaheuristics (HEMH) is presented. It combines different metaheuristics integrated with each other to enhance the search capabilities. In the proposed HEMH, the search process is divided into two phases. In the first one, the hybridization of greedy randomized adaptive search procedure (GRASP) with data mining (DM-GRASP) [106, 107] is applied to obtain an initial set of high quality solutions dispersed along the Pareto front within the framework of MOEA/D [84]. Then, the search efforts are intensified on the promising regions around these solutions through the second phase. The greedy randomized pathrelinking with local search or reproduction operators are applied to improve the quality and to guide the search

to explore the non discovered regions in the search space. The two phases are combined with a suitable evolutionary framework supporting the integration and cooperation. Moreover, the efficient solutions explored over the search are collected in an external archive. The HEMH is verified and tested against some of the state of the art MOEAs [84, 36, 14, 108] using a set of MOKSP instances used in [88] and in [84]. The experimental results indicated that the HEMH is highly competitive and can be considered as a viable alternative.

A new evolutionary clustering approach called k-mean algorithm based on multi-objective evolutionary algorithm based on decomposition (MOEA/D) [84] is developed in [109]. It optimizes two conflicting functions of data mining in its recent literature. One is snapshot quality function and the other is the history cost function. The experimental results demonstrated significantly better results than evolutionary k-mean (EKM) method.

In [110], a framework for continuous many-objective test problems with arbitrarily prescribed PS shapes is presented. Then the behavior of two popular MOEAs namely NSGA-II [14] and MOEA/D [84] are studied on the designed continuous test problems. The authors are hoped that it will promote an integrated investigation of MOEAs for their scalability with objectives and their ability to handle complicated PS shapes with varying nature of the PF.

2.4. Multimethod Search Approaches

A multialgorithm genetically adaptive for single objective optimization (AMALGAM-SO) is developed in [145]. This algorithm simultaneously combines the strengths of the covariance matrix adaptation (CMA) evolution strategy [146], genetic algorithm (GA) and particle swarm optimizer (PSO). It implements a self-adaptive learning strategy and automatically tune the number of offspring allowed to be produced by each individual algorithm based on their reproductive success at each generation. AMALGAM-SO has been tested on CEC'05 test bed of single objective optimization problems [147].

In [148], an improved version of the AMALGAMSO is developed for dealing multiobjective optimization called AMALGAM-MO. It blends the attributes of the best available individual search algorithms, NSGAII [14], PSO [111], DE [128], adaptation Metropolis search (AMS) [149]. AMALGAM tested on 2objectives ZDT test problems [44].

A novel multi-objective memetic algorithm, called multi-strategy ensemble multi-objective algorithm (MS-MOEA) is proposed in [150]. In MS-MOEA, the convergence speed is accelerated by new offspring creating operator called adaptive genetic and differential mechanism (GDM). A Gaussian mutation operator is employed to cope with premature convergence. A memory strategy is proposed for achieving better starting population when a change taken place in dynamic environment. MS-MOEA has been tested on dynamic multiobjective optimization problems.

To deal with dynamic multiobjective optimization, a new co-evolutionary algorithm (COEA) is proposed in [151]. It hybridizes competitive and cooperative mechanisms observed in universe to track the Pareto front in a dynamic environment. The main idea of the competitive-cooperative co-evolution is to allow the decomposition process of the optimization problem to adapt and emerge rather than being hand-designed and fixed at the start of the evolutionary optimization process. COEA is tested in comparison with CCEA [152], NSGA-II [14], and SPEA2 [36] on real valued test problems.

A multi-objective hybrid optimizer

denoted by MOHO is presented in [153]. MOHO combines three MOEAS, SPEA 2 [36], a multi-objective particle swarm (MOPSO) [154], and NSGA-II [14] for dealing MOPs. MOHO favors automatically the individual search algorithm that quickly improves the Pareto approximation of the MOP. MOHO grades each algorithm based on five suggested improvements criteria during its course of evolution.

In [155], the feasibility study for integration of two methods: MOEA/D [7] and NSGA-II [4] in the proposed multimethod search approach (MMTD) is performed. MMTD allocated population dynamically to both its constituent algorithms, MOEA/D [84] and NSGA-II [14], based on their individual performance during its evolutionary process. MMTD is tested on two different test suites problems, the ZDT test problems [44] and the CEC'09 test instances [48]. The final best approximated results illustrates the usefulness of MMTD dealing with multiobjective optimization (MO).

In [156], the author combined two different types MOEAs and developed a hybrid method, called MMTD. In MMTD, the whole search is divided into a number of phases. At each phase, MOEA/D and NSGA-II are run simultaneously with different computational resources based on their respective performances at the current phase of MMTD and the computational resources of the next phase are allocated to MOEA/D and NSGA-II. The effectiveness of MMTD is tested on two test suites of continuous multi-objective optimization test problems.

3. Summary

Firstly, this paper provided a general mathematical formulation to MOP and some important basic definition.

Secondly, this paper presented the literature review of some state-of-the-art hybrid evolutionary algorithms. Our literature review is organized as follows: Subsection 2.1 local search based MOEAs; Subsection 2.2 provides some hybrid versions of well-known MOEAs Based on Pareto Dominance; Subsection 2.3 includes the enhanced Versions of MOEA/D paradigm; Subsection 2.4 multi-method search approaches.

Reference:

[1] F.Y.Edgeworth, Mathematical Psychics, P. Keagan, London, England, 1881.

[2] C. A. Coello Coello, G. B.Lamont, D. A. Veldhuizen, Evolutionary Algorithms for Solving Multi-Objective Problems, Kluwer Academic Publishers, New York, 2002.

[3] K. Deb, Multi-Objective Optimization Using Evolutionary Algorithms, 2nd Edition, John Wiley and Sons Ltd, 2002.

[4] K. Deb, Multi-Objective Optimization using Evolutionary Algorithms., 1st Edition, John Wiley and Sons, Chichester, UK, 2001.

[5] K. M. Miettinien, Nonlinear Multiobjective Optimization, Kluwer's International Series, Norwell, MA: Academic Publishers Kluwer, 1999.

[6] H. Ishibuchi, T. Yoshida, T. Murata, Balance Between Genetic Search and Local Search in Memetic Algorithms for Multiobjective Permutation Flowshop Scheduling, IEEE Transactions On Evolutionary Computation 7 (2) (2003) 204–233.

[7] J. Knowles, D. Corne, Memetic Algorithms for Multiobjective Optimization: Issues, Methods and Prospects, in: a. J. S. In William E. Hart, N. Krasnogor (Ed.), Studies in Fuzziness and Soft Computing, Vol. 166, Springer, 2005, pp. 313–352.

[8] R. Thangaraj, M. Pant, A. Abraham, P. Bouvry, Particle Swarm Optimization: Hybridization Perspectives and Experimental Illustrations, Applied Mathematics and Computation 217 (12) (2011) 5208–5226.

[9] H. Ishibuchi, T. Murata., Multi-Objective Genetic Local Search Algorithm. In: I. T. Fukuda, T. Furuhashi (Eds.), Proceedings of the Third IEEE International Conference on Evolutionary Computation, 1996, pp. 119–124.

[10] H. Ishibuchi, T. Murata, Multi-Objective Genetic Local Search Algorithm and Its Application to Flowshop Scheduling, IEEE Transactions on Systems, Man and Cybernetics 28 (3) (1998) 392–403.

[11] H. Ishibuchi, T. Yoshida, T. Murata, Balance Between Genetic Search and Local Search in Memetic Algorithms for Multi-objective Permutation Flowshop Scheduling, IEEE Transactions On Evolutionary Computation 7 (2) (2004) 204–223.

[12] E. Zitzler, L. Thiele, An Evolutionary Approach for Multi-objective Optimization: The Strength Pareto Approach, TIK Report 43, Computer Engineering and Networks Laboratory (TIK), ETH Zurich (May 1998).

[13] E. Zitzler, L. Thiele, Multiobjective Evolutionary Algorithms:A comparative Case Study and the Strength Pareto Approach, IEEE Transactions On Evolutionary Computation 3 (4) (1999) 257–271.

[14] K. Deb, A. Pratap, S. Agarwal, T.Meyarivan, A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II, IEEE Transaction On Evolutionary Computation 6 (2) (2002) 182–197.

[15] A.Jaszkiewicz, On the Performance of Mutiple-Objective Genetic Local Search on the 0/1 Knapsack problems A Comparative Experiment, IEEE Transaction On Evolutionary Computation 6 (2002) 402–412.

[16] A. Jaszkiewicz, Do Multiple-Objective Metaheuristics Deliver on Their Promises? A Computational Experiment on the Set-Covering Problem., IEEE Transactions on Evolutionary Computation 7 (2) (2003) 133–143.

[17] A. Jaszkiewicz, On the Computational Efficiency of Multiple Objective Metaheuristics. The Knapsack Problem Case Study, European Journal of Operational Research 158 (2) (2004) 418– 433.

[18] H. Ishibuchi, Y. Hitotsuyanagi, N. Tsukamoto, Y. Nojima, Use of Biased Neighborhood Structures in Multiobjective Memetic Algorithms, Soft Computing -A Fusion of Foundations, Methodologies and Applications 13 (2009) 795–810.

[19] J. Knowles, D. Corne., M-PAES: A Memetic Algorithm for Multiobjective Optimization, Piscataway, New Jersey, IEEE Service Center, 2000, pp. 325–332.

[20] A. Jaszkiewicz, Genetic Local Search for Multi-Objective Combinatorial Optimization, European Journal of Operational Research 137 (1) (2002) 50–71.

[21] A. Jaszkiewicz, M. Hapke, P. Kominek, Performance of Multiple Objective Evolutionary Algorithms on a Distribution System Design Problem-Computational Experiment, in: Proceedings of First International Conference on Evolutionary Multi-Criterion Optimization (EMO), Zurich, Switzerland, 2001, pp. 241–255.

[22] J. Knowles, D. Corne, The Pareto Archived Evolution Strategy: A new Baseline Algorithm for Pareto Multiobjective Optimization, in: Proceedings of the IEEE Congress on Evolutionary Computation (CEC' 99), Piscatay, NJ, 1999, pp. 98–105.

[23] H. Wang, D. Wang, S. Yang, A Memetic Algorithm with Adaptive Hill Climbing Strategy for Dynamic Optimization Problems, Soft Comput. 13 (2009) 763–780.

[24] H. Ishibuchi, T. Doi, Y. Nojima, Incorporation of Scalarizing Fitness Functions into Evolutionary Multiobjective Optimization Algorithms, in: Lecture Notes in Computer Science, Vol. 4193: PPSN IX, Springer, 2006, pp. 493–502.

[25] E. F. Wanner, F. G. Guimar˜aes, R. H. C. Takahashi, P. J. Fleming, Local Search with Quadratic Approximations into Memetic Algorithms for Optimization with Multiple Criteria, Evol. Comput. 16 (2008) 185–224.

[26] C. Fonseca, P. Fleming, Genetic Algorithms for Multiobjective Optimization: Formulation, Discussion and Generalization, in: Proceedings of the 5th International Conference on Genetic Algorithms, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993, pp. 416–423.

[27] C. Fonseca, P. Fleming, An Overview of Evolutionary Algorithm in Multi-Objective Optimization, Evolutionary Computation 3 (1) (1995) 1–16.

[28] E. F. Wanner, F. G. Guimar˜aes, R. H. C. Takahashi, P. J. Fleming, Local Search with Quadratic Approximation in Genetic Algorithms for Expensive Optimization Problems, in: IEEE Congress on Evolutionary Computation, 2007, pp. 677–683.

[29] E. F. Wanner, F. G. Guimar˜aes, R. R. Saldanha, R. H. C. Takahashi, P. J. Fleming, Constraint quadratic approximation operator for treating equality constraints with genetic algorithms, in: Congress on Evolutionary Computation, 2005, pp. 2255–2262.

[30] E. Wanner, F. Guimaraes, R. Takahashi, D. Lowther, J. Ramirez, Multiobjective Memetic Algorithms With Quadratic Approximation-Based Local Search for Expensive Optimization in Electromagnetics, Magnetics, IEEE Transactions on 44 (6) (2008) 1126–1129.

[31] A. S. S. M. Barkat Ullah, R. Sarker, D. Cornforth, C. Lokan, An agent-based memetic algorithm (AMA) for solving constrained optimazation problems, in: IEEE Congress on Evolutionary Compu
tation CEC'7, 2007, pp. 999–1006.

[32] A. S. S. M. Barkat Ullah, R. Sarker, C. Lokan, An Agent-based Memetic Algorithm (AMA) for Nonlinear Optimization with Equality Constraints, in: IEEE Congress on Evolutionary Computation (CEC'09), 2009, pp. 70–77.

[33] A. S. S. M. Barkat Ullah, R. Sarker, D. Cornforth, C. Lokan, AMA: a New Approach for Solving Constrained Real-valued Optimization Problems, Soft Comput. 13 (2009) 741–762.

[34] A. S. S. M. B. Ullah, R. A. Sarker, D. Cornforth, A Combined MA-GA Approach for Solving Constrained Optimization Problems, in: ACIS-ICIS, 2007, pp. 382–387.

[35] A.Lara, G. Sanchez, C. Coello Coello, O.Schutze, HCS: A New Local Search Strategy for Memetic Multiobjective Evolutionary Algorithms, IEEE Transactions on Evolutionary Computation 14 (1) (2010) 112–132.

[36] E. Zitzler, M. Laumanns, L. Thiele, SPEA2: Improving the Strength Pareto Evolutionary Algorithm, TIK Report 103, Computer Engineering and Networks Laboratory (TIK), ETH Zurich, Zurich, Switzerland (2001).

[37] M. Bell, M. Pike, Remark on Algorithm 178: Direct Search, Communications of the ACM 9 (9) (1966) 684.

[38] A. Kaupe, Direct Search Algorithm 178, Communications of the ACM 6 (6) (1963) 313.

[39] R. Hooke, T. Jeeves, Direct Search Solution of Numerical and Statistical Problems, Journal of the ACM 8 (2) (1961) 212–
229.

[40] J. Fliege, L. M. G. n. Drummond, B. F. Svaiter, Newton's Method for Multiobjective Optimization, SIAM J. on Optimization 20 (2009) 602–626.

[41] B.T.Polyak, Newton's Method and its use in Optimization, European Journal of Operational Research 181 (3) (2007) 1086– 1096.

[42] N. Beume, B. Naujoks, M. Emmerich, SMS-EMOA: Multiobjective Selection based on Dominated hypervolume, European Journal of Operational Research 181 (3) (2007) 1653–1669.

[43] P. Koch, O. Kramer, G. Rudolph, B. Nicola, On the hybridization of SMS-EMOA and local search for continuous multi-objective optimization, in: Proceedings of the 11th Annual conference on Genetic and evolutionary computation, GECCO '09, ACM, New York, NY, USA, 2009, pp. 603–610.

[44] E. Zitzler, K. Deb, L. Thiele, Comparsion of Multiobjective Evolutionary Algorithms: Emperical Results, Evolutionary Computation 8 (2) (200) 173–195.

[45] L. Y. Tseng, C. Chen, Multiple Trajectory Search for Multiobjective Optimization, in: Proceedings of the Congress on Evolutionary ComputationEvolutionary Computation, CEC'07, IEEE Press, Singapore, 2007, pp. 3609–3616.

[46] V. L. Huang, A. K. Qin, K. Deb, E. Zitzler, P. N. Suganthan, J. J. Liang, M. Preuss, S. Huband, Problem Definitions for Performance Assessment of Multi-Objective Optimization Algorithms,

Technical report, Nanyang Technological University, Singapore (25–28 September, 2007).

[47] L. Y. Tseng, C. Chen, Multiple trajectory search for unconstrained/constrained multi-objective optimization, in: Proceedings of the Eleventh conference on Congress on Evolutionary Computation, CEC'09, IEEE Press, Piscataway, NJ, USA, 2009, pp. 1951–1958.

[48] Q. Zhang, A. Zhou, S. Zhaoy, P. N. Suganthany, W. Liu, S. Tiwariz, Multiobjective Optimization Test Instances for the CEC 2009 Special Session and Competition, Technical Report CES487 (2009).

[49] M. Gong, C. Liu, L. Jiao, G. Cheng, Hybrid Immune Algorithm with Lamarckian Local Search for Multi-Objective Optimization, Memetic Computing 2 (2010) 47–67.

[50] M. Gong, L. Jiao, H. Du, L. Bo, Multiobjective Immune Algorithm with Nondominated Neighbor-Based Selection, Evolutionary Computation 16 (2) (2008) 225–255.

[51] C. Grosan, A. Abraham, Approximating Pareto Frontier Using a Hybrid Line Search Approach, Inf. Sci. 180 (14) (2010) 2674–2695.

[52] T. Okabe, Evolutionary Multi-Objective Optimization-On the Distribution of Offspring in Parameter and Fitness Space, PhD thesis, Bielefeld University, Germany (2004).

[53] K. Deb, L. Thiele, M. Laumanns, E. Zitzler, Scalable MultiObjective Optimization Test Problems, In Congress on Evolutionary Computation (CEC2002), Piscataway, New Jersey: IEEE service Center 1 (2002) 825–830.

[54] P. K. Tripathi, S. Bandyopadhyay, S. K. Pal, Multi-Objective Particle Swarm Optimization with Time Variant Inertia and Acceleration Coefficients, Information Science 177 (22) (2007) 5033–5049.

[55] H. Ishibuchi, T. Yoshida, Hybrid Evolutionary Multi-Objective Optimization Algorithms, in: Design, Management and Applications, Frontiers in Artificial Intelligence and Applications, IOS Press, 2002, pp. 163–172.

[56] T. Murata, S. Kaige, H. Ishibuchi, Generalization of Dominance Relation-Based Replacement Rules for Memetic EMO Algorithms, in: Proc. of 2003 Genetic and Evolutionary Computation Conference, 2003, pp. 1234–1245.

[57] S. F. Adra, T. J. Dodd, I. A. Griffin, P. J. Fleming, Convergence Acceleration Operator for Multiobjective Optimization, Trans. Evol. Comp 13 (2009) 825–847.

[58] S. F. Adra, I. Griffin, P. J. Fleming, Hybrid Multiobjective Genetic Algorithm with a New adaptive Local Search Process, Proceedings of the 2005 conference on Genetic and evolutionary computation GECCO 05 1 1009–1010.

[59] D. Sharma, A. Kumar, K. Deb, K. Sindha, Hybridization of SBX based NSGA-II and Sequential Quadratic Programming for solving Multi-objective Optimization Problems, Technical Report 2007007, Kanpur Genetic Algorithms Laboratory (KanGAL), India (2009).

[60] D. Sharma, A. Kumar, K. Deb, K. Sindhya, Hybridization of SBX based NSGA-II and Sequential Quadratic Programming for Solving Multi-Objective Optimization Problems, in: IEEE Congress on Evolutionary Computation, CEC'07, 2007, pp. 3003–3010.

[61] A. Kumar, D. Sharma, K. Deb, A Hybrid Multi-Objective Optimization Procedure using PCX Based NSGA-II and Sequential Quadratic Programming, in: Proceedings of the IEEE Congress on Evolutionary Computation, CEC'07, 25—28 September 2007, Singapore, 2007, pp. 3011—3018.

[62] K. Deb, K. Miettinen, S. Chaudhuri, Toward an Estimation of Nadir Objective Vector Using a Hybrid of Evolutionary and Local Search Approaches, IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, 14 (6) (2010) 821–841.

[63] K. Sindhya, K. Deb, K. Miettinen, A Local Search Based Evolutionary Multi-Objective Optimization Approach for Fast and Accurate Convergence, in: Proceedings of the 10th international conference on Parallel Problem Solving from Nature: PPSN X, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 815– 824.

[64] A.P.Wierzbicki, The Use of Reference Objectives in Multiobjective Optimization., in: F. G., G. T. (Eds.), MCDM theory and Application, Proceedings, no. 177 in Lecture Notes in Eco-

nomics and Mathematical Systems, Springer Verlag, Hagen, 1980, pp. 468–486.

[65] K. Sindhya, A. Sinha, K. Deb, K. Miettinen, Local Search Based Evolutionary Multi-Objective Optimization Algorithm for Constrained and Unconstrained Problems, in: Proceedings of the Eleventh conference on Congress on Evolutionary Computation, CEC'09, IEEE Press, Piscataway, NJ, USA, 2009, pp. 2919–2926.

[66] Y. Wang, Z. Cai, G. Guo, Y. Zhou, Multi-objective Optimization and Hybrid Evolutionary Algorithm to Solve Constrained Optimization Problems, IEEE Transactions On Systems,Man and Cybernetics-PartB:Cybernetics 37 (3) (2007) 560–575.

[67] P. Koduru, Z. Dong, S. Das, S. Welch, J. Roe, E. Charbit, A Multiobjective Evolutionary-Simplex Hybrid Approach for the Optimization of Differential Equation Models of Gene Networks, Evolutionary Computation, IEEE Transactions on 12 (5) (2008) 572-590.

[68] J. Nelder, R. Mead., A Simplex Method for Functions Minimizations, Computer Journal 7 (4) (1965) 308–313.

[69] A. K. M. K. A. Talukder, M. Kirley, R. Buyya, A Pareto Following Variation Operator for Fast-Converging Multiobjective Evolutionary Algorithms, in: GECCO, 2008, pp. 721–728.

[70] Q. Zhang, A. Zhou, Y. Jin, RM-MEDA: A Regularity Model Based Multiobjective Estimation of Distribution Algorithm, IEEE Transactions On Evolutionary Computation 12 (1) (2008) 41–63.

[71] A. K. M. Khaled Ahsan Talukder, M. Kirley, R. Buyya, The Pareto-Following Variation Operator as an Alternative Approximation Model, in: IEEE Congress on Evolutionary Computation, CEC 2009, Trondheim, Norway, 18-21 May, 2009, 2009, pp. 8–15.

[72] S. Tiwari, P. Koch, G. Fadel, K. Deb, AMGA: an archive-based micro genetic algorithm for multi-objective optimization, in: GECCO, 2008, pp. 729–736.

[73] S. Tiwari, G. Fadel, P. Koch, K. Deb, Performance assessment of the hybrid archive-based micro genetic algorithm (AMGA) on the CEC09 test problems, in: Proceedings of the Eleventh conference on Congress on Evolutionary Computation, CEC'09, IEEE Press, Piscataway, NJ, USA, 2009, pp. 1935-1942.

[74] N. Hamada, J. Sakuma, S. Kobayashi, I. Ono, and Functional-Specialization Multi-Objective Real-Coded Genetic Algorithm: FS-MOGA, in: G. Rudolph, T. Jansen, S. Lucas,
C. Poloni, N. Beume (Eds.), Parallel Problem Solving from Nature PPSN X, Vol. 5199 of Lecture Notes in Computer Science, Springer Berlin /Heidelberg, 2008, pp. 691–701.

[75] K. Deb, T. Goel, A Hybrid Multi-objective Evolutionary Approach to Engineering Shape Design, in: Proceedings of the First International Conference on Evolutionary MultiCriterion Optimization (EMO'01), Zurich, Switzerland, 2001, pp. 385–
399.

[76] A. Hernan, T. Kiyoshi, A Hybrid Selection Strategy Using Scalarization and Adaptive epsilon-Ranking for Many-objective Optimization, Transaction of the Japanese Society for Evolutionary Computation (TJNSEC) 1 (1) (2010) 65-78.

[77] H. Aguirre, K. Tanaka, A Hybrid Scalarization and Adaptive ϱ-Ranking Strategy for Many-Objective Optimization, in: R. Schaefer, C. Cotta, J. Kolodziej, G. Rudolph (Eds.), Parallel Problem Solving from Nature PPSN XI, Vol. 6239 of Lecture Notes in Computer Science, Springer Berlin /Heidelberg, 2011, pp. 11–20.

[78] S. Z. Mart´ınez, C. A. Coello Coello, A Proposal to Hybridize Multi-Objective Evolutionary Algorithms with Non-Gradient Mathematical Programming Techniques, in: Proceedings of-Parallel Problem Solving from Nature -PPSN X, 10th International Conference Dortmund, Germany, September 13-17, 2008, pp. 837–846.

[79] S. Zapotecas Mart´ınez, C. A. Coello Coello, Hybridizing an Evolutionary Algorithm with Mathematical Programming Techniques for Multi-Objective Optimization, in: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation, GECCO '08, ACM, New York, NY, USA, 2008, pp. 769–770.

[80] Y. Yang, J.-f. Wu, X.-b. Zhu, J.-c. Wu, A Hybrid Evolutionary Algorithm for finding Pareto optimal set in Multi-Objective Optimization, in: 2011 Seventh International Conference on Natural Computation (ICNC), Vol. 3, 2011, pp. 1233–1236.

[81] J.Knowles, Parego:A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems, IEEE Transactions on Evolutionary Computation 10 (1) (2006) 50–66.

[82] J.Knowles, ParEGO: a Hybrid Algorithm with on-line Landscape Approximation For Expensive Multiobjective Optimization Problems, IEEE Transactions on Evolutionary Computation, 10 (1) (2006) 50–66. doi:10.1109/TEVC.2005.851274.

[83] D. R. Jones, M. Schonlau, W. J. Welch, Efficient Global Optimization of Expensive Black-Box Functions, Journal of Global Optimization 13 (1998) 455–492.

[84] Q. Zhang, H. Li, MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition , IEEE transaction on Evolutionary Computation 11 (6) (2007) 712–731.

[85] W. Peng, Q. Zhang, H. Li, Comparison between MOEA/D and NSGA-II on the Multiobjective Travelling Salesman Problem, Technical Report CES-478, Department of Computing and Electronic Systems University of Essex (December 2007).

[86] H. Li, Q. Zhang, Multiobjective Optimization Problems With Complicated Pareto Sets: MOEA/D and NSGA-II, IEEE Transsation On Evolutionary Computation 13 (2) (2009) 284–302.

[87] P. Palmers, T. McConaghy, M. Steyaert, G. G. E. Gielen, Massively multi-topology sizing of analog integrated circuits, in: Design, Automation and Test in Europe (DATE), 2009, pp. 706–711.

[88] H. Ishibuchi, Y. Sakane, N. Tsukamoto, Y. Nojima, Adaptation of Scalarizing Functions in MOEA/D: An Adaptive Scalarizing Function-Based Multiobjective Evolutionary Algorithm, in: Proceedings of the 5th International Conference on Evolutionary Multi-Criterion Optimization, EMO '09, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 438–452.

[89] Q. Zhang, W. Liu, E. Tsang, B. Virginas, Expensive multiobjective optimization by MOEA/D with Gaussian process model, Trans. Evol. Comp 14 (2010) 456–474.

[90] W. Liu, Q. Zhang, E. Tsang, C. Liu, B. Virginas, On the performance of metamodel assisted MOEA/D, in: Proceedings of the 2nd international conference on Advances in computation and intelligence, ISICA'07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 547–557.

[91] H. Li, J. Landa-Silva, Evolutionary Multi-objective Simulated Annealing with Adaptive and Competitive Search Direction, in: Proceedings of IEEE Congress on Evolutionary Computation (CEC'08), Hong Kong, 2008, pp. 3310–3317.

[92] Q. Zhang, H. Li, D. Maringer, E. P. K. Tsang, MOEA/D with NBI-style Tchebycheff Approach for Portfolio Management, in: Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2010, Barcelona, Spain, 18-23 July 2010, 2010, pp. 1–8.

[93] C.-M. Chen, Y.-P. Chen, Q. Zhang, Enhancing MOEA/D with Guided Mutation and Priority Update for Multi-Objective Optimization, in: Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2009, Trondheim, Norway, 18-21 May, 2009, 2009, pp. 209–216.

[94] N. Al Moubayed, A. Petrovski, J. McCall, A novel smart multi-objective particle swarm optimisation using decomposition, in: Proceedings of the 11th international conference on Parallel problem solving from nature: Part II, PPSN'10, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 1–10.

[95] B. Liu, F. V. Fern´andez, Q. Zhang, M. Pak, S. Sipahi, G. G. E. Gielen, An Enhanced MOEA/D-DE and its Application to Multiobjective Analog Cell Sizing, in: IEEE Congress on Evolutionary Computation, 2010, pp. 1–7.

[96] M. Nasir, A. K. Mondal, S. Sengupta, S. Das, A. Abraham, An Improved Multiobjective Evolutionary Algorithm based on Decomposition with Fuzzy Dominance, in: Proceedings of IEEE Congress on Evolutionary Computation (CEC,11), IEEE Press, New Orleans, US, 2011, pp. 1–8.

[97] M. Gong, F. Liu, W. Zhang, L. Jiao, Q. Zhang, Interactive MOEA/D for Multi-Objective Decision Making, in: Proceedings of 13th Annual Genetic and Evolutionary Computation Conference, GECCO 2011, Dublin, Ireland, July

12-16, 2011, 2011, pp. 721–728.

[98] R. Battiti, A. Passerini, Brain-Computer Evolutionary Multiobjective Optimization: A Genetic Algorithm Adapting to the Decision Maker, IEEE Trans. Evolutionary Computation 14 (5) (2010) 671–687.

[99] H. Ishibuchi, Y. Hitotsuyanagi, H. Ohyanagi, Y. Nojima, Effects of the Existence of Highly Correlated Objectives on the Behavior of MOEA/D, in: Proceedings of the 6th international conference on Evolutionary multi-criterion optimization, EMO'11, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 166– 181.

[100] S. Jiang, Z. Cai, J. Zhang, Y.-S. Ong, Multiobjective Optimization by Decomposition with Pareto-adaptive Weight Vectors, in: Seventh International Conference on Natural Computation, ICNC 2011, Shanghai, China, 26-28 July, 2011, 2011, pp. 1260–1264.

[101] W. Khan, Q. Zhang, MOEA/D-DRA with Two Crossover Operators, in: Proceedings of the UK Workshop on Computational Intelligence (UKCI 2010), 2010, pp. 1–6.

[102] Q. Zhang, W. Liu, H. Li, The Performance of a New Version of MOEA/D on CEC'09 Unconstrained MOP Test Instances, IEEE Congress On Evolutionary Computation (IEEE CEC 2009), Trondheim, Norway (2009) 203-208.

[103] Y. Mei, K. Tang, X. Yao, Decomposition-Based Memetic Algorithm for Multiobjective Capacitated Arc Routing Problem, IEEE Trans. Evolutionary Computation 15 (2) (2011) 151–165.

[104] P. Lacomme, C. Prins, M. Sevaux, A Genetic Algorithm for biobjective Capacitated Arc Routing Problem, Computer and operations Research 33 (12) (2006) 3473–3493.

[105] A.Kafafy, A. Bounekkar, S. Bonnevay, A Hybrid Evolutionary Metaheuristics (HEMH) applied on 0/1 Multiobjective Knapsack Problems, in: Proceedings of the 13th annual conference on Genetic and evolutionary computation, GECCO '11, ACM, New York, NY, USA, 2011, pp. 497–504.

[106] M. H. Ribeiro, A. Plastino, S. L. Martins, Hybridization of GRASP Metaheuristic with Data Mining Techniques, J. Math. Model. Algorithms 5 (1) (2006) 23–41.

[107] Santos, F. Luis, Martins, L. Simone, Plastino, Alexandre, Applications of the DM-GRASP heuristic: A Survey, International Transactions in Operational Research 15 (4) (2008) 387– 416.

[108] D. S. Vianna, J. E. C. Arroyo, A GRASP Algorithm for the Multi-Objective Knapsack Problem, in: XXIV International Conference of the Chilean Computer Science Society (SCCC 2004), 11-12 November 2004, Arica, Chile, 2004, pp. 69–75.

[109] J. Ma, Y. Wang, M. Gong, L. Jiao, Q. Zhang, Spatio-Temporal Data Evolutionary Clustering Based on MOEA/D, in: Proceedings of the 13th annual conference companion on Genetic and evolutionary computation, GECCO '11, ACM, New York, NY, USA, 2011, pp. 85–86.

[110] D. K. Saxena, Q. Zhang, J. A. Duro, A. Tiwari, Framework for Many-Objective Test Problems with Both Simple and Complicated Pareto-Set Shapes, in: Proceedings of Evolutionary Multiobjective Optimization, 2011, pp. 197–211.

[111] R. Eberhart, J.Kennedy, A New Optimizer Using Particle Swarm Theory, in: Proceedings of the Sixth International Symposium on Micro Machine and Human Science, MHS'95, 1995, pp. 39–43.

[112] M. Reyes-Sierra, C. A.Colleo Colleo, Multi-Objective particle Swarm Optimizers:A Survey of the State-of-Aart, International Journal of Computation Intelligence Reaserch. 2 (3) (2006) 287–308.

[113] A. Banks, J.Vincent, C. Anyakoha, A Review of Particle Swarm optimization. Part I: Background and Development 6 (2007) 467–484.

[114] A.Banks, J.Vincent, C. Anyakoha, A Review of Particle Swarm Optimization. Part II: Hybridisation, Combinatorial, Multicriteria and Constrained Optimization, And Indicative Applications 7 (2008) 109–124.

[115] Y. Shi, R. Eberhart, A Modified Particle Swarm Optimizer, in: Proceedings of IEEE International Conference on Evolutionary Computation, CEC'98, 1998, pp. 69–73.

[116] C. S. Tsou, S. C. Chang, P. W. Lai, Using

Crowding Distance to Improve Multi-Objective PSO with Local SearchSource: Swarm Intelligence: Focus on Ant and Particle Swarm Optimization, Book edited by: Felix T. S. Chan and Manoj Kumar Tiwari, ISBN 978-3-902613-09-7, pp. 532, December 2007, Itech Education and Publishing, Vienna, Austria.

[117] C. R. Raquel, J. Prospero C. Naval, An Effective use of Crowding Distance in Multiobjective Particle Swarm Optimization, in: Proceedings of the 2005 conference on Genetic and evolutionary computation, GECCO '05, ACM, New York, NY, USA, 2005, pp. 257–264.

[118] L. Santana-Quintero, N. Ramrez-Santiago, C. Coello, J. Luque, A. Hernndez-Daz, A New Proposal for Multiobjective Optimization Using Particle Swarm Optimization and Rough Sets Theory, in: T. Runarsson, H.-G. Beyer, E. Burke, J. Merelo-Guervs, L. Whitley, X. Yao (Eds.), Parallel Problem Solving from Nature - PPSN IX, Vol. 4193 of Lecture Notes in Computer Science, Springer Berlin /Heidelberg, 2006, pp. 483–492.

[119] Z.Pawlak, Rough Sets–Theoretical Aspects of Reasoning about Data, Kluwer Academic, Dordrecht, Netherland, 1991.

[120] A. Elhossini, S. Areibi, R. Dony, Strength Pareto Particle Swarm Optimization and Hybrid EA-PSO for Multi-Objective Optimization, Evol. Comput. 18 (2010) 127–156.

[121] C. A. C. Coello, G. T. Pulido, M. S. Lechuga, Handling Multiple Objectives with Particle Swarm Optimization, IEEE Transactions on Evolutionary Computation 8 (2004) 256–279.

[122] L. C. Cagnina, S. C. Esquivel, Solving Hard Multiobjective Problems with a Hybridized Method, Journal of Computer Science and Technology (JCS & T) 10 (3) (2010) 117–122.

[123] L. Cagnina, S. Esquivel, C. Coello, A bi-Population PSO with a Shake-Mechanism for Solving Constrained Numerical Optimization, in: IEEE Congress on Evolutionary Computation, 2007. CEC 2007., 2007, pp. 670–676.

[124] X. Wang, L. Tang, A PSO-Based Hybrid Multi-Objective Algorithm for Multi-Objective Optimization Problems, in: Y. Tan, Y. Shi, Y. Chai, G. Wang (Eds.), Advances in Swarm Intelligence, Vol. 6729 of Lecture Notes in Computer Science, Springer Berlin /Heidelberg, 2011, pp. 26–33.

[125] M. Lashkargir, S. A. Monadjemi, A. Baraani-dastjerdi, A Hybrid Multi Objective Particle Swarm Optimization Method to Discover Biclusters in Microarray Data, (International Journal of Computer Science and Information Security(IJCSIS) 4 (1).

[126] L. Benameur, J. Alami, A. El Imrani, A New Hybrid Particle Swarm Optimization Algorithm for Handling Multiobjective Problem Using Fuzzy Clustering Technique, in: International Conference on Computational Intelligence, Modelling and Simulation, CSSim '09., 2009, pp. 48–53.

[127] R.Storn, K.V.Price, Differential Evolution -A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces, Technical Report TR-95-012, ICSI (1995).

[128] R.Storn, K.V.Price, Differential Evolution -a Simple and Efficient Heuristic for Global Optimization over Continuous Spaces, J.Global Opt 11 (4) (1997) 341–359.

[129] F. Neri, V. Tirronen, Recent Advances in Differential Evolution: A Survey and Experimental Analysis, Artificial Intelligence Review 33 (2010) 61–106. [130] S. Das, P. N. Suganthan, Differential Evolution: A Survey of the state-of-the-art, IEEE Trans. Evolutionary Computation 15 (1) (2011) 4–31.

[131] U. K. Chakraborty, Advances in Differential Evolution, Springer Publishing Company, Incorporated, 2008.

[132] M. G. Epitropakis, D. K. Tasoulis, N. G. Pavlidis, V. P. Plagianakos, M. N. Vrahatis, Enhancing Differential Evolution Utilizing Proximity-Based Mutation Operators, IEEE Trans. Evolutionary Computation 15 (1) (2011) 99–119.

[133] H.-Y. FAN, J. LAMPINEN, A Trigonometric Mutation Operation to Differential Evolution, Journal of Global Optimization 27 (2003) 105–129.

[134] N. Pavlidis, V. Plagianakos, D. Tasoulis, M.

Vrahatis, Human Designed Vs. Genetically Programmed Differential Evolution Operators, in: Proceeding of IEEE Congress on Evolutionary Computation, IEEE Press, heraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada, 2006, pp. 1880–1886.

[135] J. J. Durillo, A. J. Nebro, F. Luna, E. Alba, Solving Three-Objective Optimization Problems Using a New Hybrid Cellular Genetic Algorithm, in: Proceedings of the 10th international conference on Parallel Problem Solving from Nature: PPSN X, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 661– 670.

[136] A. J. Nebro, J. J. Durillo, F. Luna, B. Dorronsoro, E. Alba, MOCell: A Cellular Genetic Algorithm for Multiobjective Optimization, International Journal of Intelligent Systems (2007) 25–36.

[137] S. Kukkonen, J. Lampinen, GDE3: the Third Evolution step of Generalized Differential Evolution, in: IEEE Congress on Evolutionary Computation, 2005, pp. 443–450.

[138] A. Zamuda, J. Brest, B. Boskovic, V. Zumer, Differential Evolution with Self-adaptation and Local Search for Constrained Multiobjective Optimization, in: IEEE Congress on Evolutionary Computation, CEC'09, Trondheim, Norway, 18-21 May, 2009, pp. 195–202.

[139] A. Zamuda, J. Brest, B. Boskovic, V. Zumer, Differential evolution for multiobjective optimization with self adaptation, in: IEEE Congress on Evolutionary Computation, 2007, pp. 3617– 3624.

[140] J. Brest, S. Greiner, B. Boskovic, M. Mernik, V. Zumer, Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems, IEEE Trans. Evolutionary Computation 10 (6) (2006) 646–657.

[141] R. L. Becerra, C. A. C. Coello, Epsilon-Constraint with an Efficient Cultured Differential Evolution, in: GECCO (Companion), 2007, pp. 2787-2794.

[142] S.Huband, P.Hingston, L. L.While, A Review of Multiobjective Test Problems and a Scalable Test Problem Toolkit, IEEE Transactions on Evolutionary Computation 10 (5) (2006) 477–506.

[143] R. Reynolds, An Introduction to Cultural Algorithms, in: R. Reynolds, L. Fogel (Eds.), Proceeding of 3rd Annual Conference on Evolutionary Programming, 1994, pp. 131–139.

[144] A. G. Hern´andez-D´ıaz, L. V. Santana-Quintero, C. Coello Coello, R. Caballero, J. Molina, A New Proposal for Multi-Objective Optimization using Differential Evolution and Rough Sets Theory, in: Proceedings of the 8th annual conference on Genetic and evolutionary computation, GECCO '06, ACM, New York, NY, USA, 2006, pp. 675–682.

[145] J. A. Vrugt, B. A. Robinson, J. M. Hyman, Self-Adapotive Multimethod Search for Global Optimization in Real-Parameter Spaces, IEEE Transsation On Evolutionary Computation 13 (2) (2009) 243–259.

[146] N. Hansen, S. Kern, Evaluating the CMA Evolution Strategy on Multimodal Test Functions, in: Proceedings of International Conference on Parallel Problem Solving from Nature, PPSN VIII, 2004, pp. 282–291.

[147] P.N.Suganthan, N. J.J.liang, K.Deb, Y.P.Chen, A. Auger, S.Tiwari, Problem Definition and Evaluation Criteria for the CEC 2005 Special Ssession on Real-Optimization, Technical Report, Nayan Technology University, Singapore and Kangal Report Number 2005005(Kanpur Genetic Algorithnms Laboratory, IIT Kanpur) (May 2005).

[148] J. A. Vrugt, B. A. Robinson, Improved Evolutionary Optimization from genetically adaptive mutimethod search, Proceedings of the National Academy of Sciences of the United States of America: PNAS (USA) 104 (3) (2007) 708–701.

[149] H.Haario, E.Saksman, J. Tamminen, An adaptive Metropolis Algorithm, Official Jounral of Bernoulli Society for Mathematical Statistics and Probability 7 (2) (2001) 223–242.

[150] Y. Wang, B. Li, Multi-Strategy Ensemble Evolutionary Algorithm for Dynamic Multi-Objective Optimization, Memetic Computing 2 (2010) 3–24.

[151] C. K. Goh, K. C. Tan, A Competitive-Cooperative Coevolutionary Paradigm for

Dynamic Multiobjective Optimization, Trans. Evol. Comp 13 (2009) 103–127.

[152] K. C. Tan, Y. J. Yang, C. K. Goh, A Distributed Cooperative Co-evolutionary Algorithm for Multiobjective Optimization, IEEE Trans. Evolutionary Computation 10 (5) (2006) 527–549.

[153] R. J. Moral, G. S. Dulikravich, Multi-Objective Hybrid Evolutionary Optimization with Automatic Switching Among Constituent Algorithms, AIAA Journal 46 (3) (2008) 673–681.

[154] K. E. Parsopoulos, M. N. Vrahatis, Particle Swarm Optimization Method in Multiobjective Problems, in: In Proceedings of the ACM Symposium on Applied Computing (SAC'02, ACM Press, 2002, pp. 603–607.

[155] W. K. Mashwani, Integration of NSGA-II and MOEA/D in Multimethod Search Approach, in: Proceedings of the 13th annual conference companion on Genetic and evolutionary computation, GECCO '11, ACM, New York, NY, USA, 2011, pp. 75–76.

[156] W. K. Mashwani, A Multimethod Search Approach Based on Adaptive Generations Level, in: Seventh International Conference on Natural Computation, ICNC 2011, Shanghai, China, 26-28 July, 2011, 2011, pp. 23-27.

Reengineering multi tiered enterprise business applications for performance enhancement and reciprocal or rectangular hyperbolic relation of variation of data transportation time with row pre-fetch size of relational database drivers

[1]Sridhar Sowmiyanarayanan

[1]Technology Excellence Group, Banking and Financial Services Technologies, Tata Consultancy Service Ltd, Bangalore, Karnataka, India

Abstract

In a traditional multitier applications performance bottlenecks can be in user interactions level or network latency or data access or business logic level. The solutions as changes or tuning parameters can be applied at architect, design, framework or algorithm or at coding level. This paper highlights an inquisitive, experimental, top down, tear apart, drill down and analytical approach across two aspects one across end to end process flow and data flow on those specific use cases or scenarios requiring performance improvement and another across layers of abstraction like architecture, framework, design, logic and coding. Re engineering for performance gain requires identifying hot spots on both aspects viz which architectural, design decision or which processing or data flow stage is having performance issue. Once identified one can further drill down and identify root cause and also can find solution as a change. To help the application owner in decision making process, the analysis outcome should have tuning parameters, relationship between them, optimum values, tradeoffs on each changes, effort, risks, cost and benefits for incorporating each change. Following the above mentioned approach on a system in production with large enterprise we could drill down to a rectangular hyperbolic or reciprocal relation between elapsed time to transport all records retrieved from a query and the number of records being pre fetched (pre fetch size) and cached in the data base client application by the database driver in each trip. Because of the reciprocal nature , we could observe that when the pre fetch size is low drastic reduction in elapsed time could be obtained even for a small increase in pre fetch size, whereas when the pre fetch size is high the gain in performance is not so significantly high even for larger increase in pre fetch size.

Keywords: *Pre fetch size, JDBC, Query result set, rectangular hyperbolic relation, data transportation time, network hops.*

Introduction

There is growing need and challenges to re engineer existing business applications in production to improve on quality attributes like performance and scalability. Business applications systems built without focusing on non functional requirements or quality attributes and their future growth in demand are facing the need and challenges of reengineering for improving quality attributes. Some quality attributes like reliability requirement may not change or increase over a period of time but performance and scalability demand may increase due to increasing number of concurrent users of the system or increasing data volume which needs to be processed by the system. Among a large set of different business applications though there may be many reasons for performance bottlenecks in traditional n tier system, the most common performance bottleneck area could be in data access layer or avoidable high memory footprint of the application.

A general approach to re engineer for performance improvement is to elicit and extract the implemented architect and design. Understand the various stages of process and data flow focusing on those scenarios or use cases for which performance gain is required. One can time profile across various processing and data flow stages to identify hot spots and can further drill down in to details with experimentation and measurements to identify root cause. Once root cause is known solutions can

be found as changes. Changes can be at various levels of abstractions like at architecture, design, logic etc. Estimates of performance gain, effort required for incorporating each change, tradeoffs for each changes will help the application owner to decide on cost benefit and other impacts to incorporate each recommended changes. Such analytical or experimental relations between tuning parameters will be greatly helpful because such relations need not be specific to particular application but can be generic and same analytical relations can be reused for tuning similar applications with similar environment in which such relations are valid.

Re engineering for performance gain and quantitative analysis both experimental and analytical on data transportation between application server and database done on a java j2ee banking application for bank's customers to view online reports related to their assets under banks custody is detailed here as a case study application. The outcome of the case study as a reciprocal relationship between total time (T) to transport query results of N records and the pre fetch size (f) is discussed. Though the relation between T and f may have other factors, we could approximate to the reciprocal relation and rectangular hyperbolic trend as a dominant factor.

1. Need for performance enhancements

Many enterprises had over the years built software applications to meet their operational, transactional or for customer services. Over the period of time, the load in terms of number of users or size of data being processed had grown but the built system had not scaled proportionately. This resulted in low performance of the application and or its inability to scale to meet the increasing demand. Some of the reasons any enterprise in general or banking and financial sector to opt for performance or scalability enhancements of their applications are listed below.

1. The technical and application architecture would not have been planned appropriately taking in to account the non functional requirements properly as the development team would have focused primarily on functional requirements. Other common reason could be that the system might have evolved over few years in ad hoc manner rather than planned, architected, well designed and built.
2. The load in terms of number of users or number of files or number data records to process would have increased recently but the system could not meet new increased load.
3. The load in terms of size of file and or size of messages to process would have increased recently.
4. Additional functionalities were or needed to be rolled out but the system could not perform or scale to the additional functional needs.

2. Common Non Functional Requirements which can change with time with higher demand

For many business applications, the demand for some quality attributes or non functional requirements like reliability may not change over a period of time whereas demand for other quality attributes like performance may increase with time. The three most common qualities of service, the demand for which can increase over a period of time with some illustrative examples are

2.1 Performance

a. User response time in the case of interactive applications or

b. Processing time in case of batch (non interactive) applications.

2.2 Scalability

2.1.1. Number of online concurrent users to support in the case of interactive applications.

2.1.2. Number of records or files or messages to process in the case of batch applications.

2.1.3. Size of file or size of message being processed.

2.2. Throughput

For example number of work items completed over period of time. Work items for example can be number of trades to be prized or number of trades to be settled or number of transactions to be completed or number of pairs of records to be reconciled.

3. Possible performance problem areas

In traditional N tier client application server database server enterprise business applications, performance bottlenecks can be in any or more of

3.1. Inefficient data access from database

3.2. Inefficient data transport between client and application server or application server and database

3.3. Inefficient I/O operations like file I/Os

3.4. Network latency

3.5. Inefficient application logic and inefficient algorithms

3.6. Resource contention like contention for CPU and memory as clients or user sessions, server executables or threads waiting for their turn for these resources

3.7. Higher memory foot print or frequent memory page faults resulting in higher percentage of disk based I/O.

Above are not comprehensive but common performance bottlenecks. Among these inefficient data access and inefficient data transport from database are the most common cause of performance bottlenecks in many applications.

4. General approaches to reengineer for performance optimization.

Assume that a banking enterprise engages a software services vendor to enhance the performance of its existing web client-application server-data base tiered application in production. Though there can be multiple ways and approaches any vendor can adopt, following describes one possible approach and steps to reengineer the application to improve on performance.

4.1. Elicit and enumerate the list of scenarios and use cases which are having low quality of services such as low performance and get a scope of problem areas.

4.2. Elicit and understand the implemented architecture, high level design, high level data and process flow and get a high level bird's eye view of functional, technical, process and data flow overview.

4.3. Breakup the high level flow in to various smaller processing stages. Instrument the code or use profilers and tools and do test run to record elapsed time breakups across various stages of the application flow.

4.4. Identify where maximum time is spent across various stages.

4.5. Identify list of root causes for the performance bottlenecks within identified low performing stages

4.6. Find solution as changes to rectify the root causes.

4.7. Changes can range from simple parameter changes to changes at code, logic design and at architecture level.

4.8. Measure or estimate benefits for each parameter changes and get a quantitative and analytical detail on each parameter.

4.9. Parameters don't impact the performance in isolations and changes in one parameter may influence the benefit derived due to changes made in other parameters.

4.10. For each change identified as a solution to improve performance, estimate quantitatively or qualitatively the following

4.10.1. Tradeoffs

4.10.2. Cost of applying the changes and benefits if the changes are incorporated

4.10.3. Risks associated with each change.

4.10.4. For example using memory caching to store all application data instead of persistent database can be a suggested change to improve performance, but the trade off are

4.10.4.1. Between performance and scalability to high data volume.

4.10.4.2. Licensing cost of the caching product and cost for the effort to change and test.

4.10.4.3. Risk of reliability in terms of data loss if there are system crashes/failures and if there are no failover recovery mechanism implemented for data in volatile memory.

5. Various levels of performance problems and solutions

The problem and solution can be at various levels of abstraction as given below.

Table 1: Illustrative performance problems at various levels and sample solution of a typical multi tiered application

Levels of abstraction	**Example problem areas**	**Example change scenarios to gain performance**
Architecture level	Disk based I/O	Caching can be used to reduce latency.
	Application access data from remote locations	Deployment architecture can be changed to co-locate application and data in same geography to avoid networked data access.
	User responses found to slow down, when large number of simultaneous users made request or processing time slows down when large number of data records has to be processed.	Request processing can be parallelized by more server instances and load balancing. If processing of one data record is independent of other data record, then data records processing can be parallelized by many server instances.
Design level	I/O and CPU are sequential though there is opportunity to do both concurrently.	Threads can be used to concurrently do I/O and processing
Framework level	Framework has too many layers: Assume that the used framework introduced too many layers of indirection and data conversion and transportation. For example used framework marshals and un marshal's data across network, converts data from flat file to XML to java object to data base objects using ORM (object relational mapping software).	More appropriate framework can be chosen to avoid or minimize data marshalling, un marshalling, conversions and transportations.
Logic level	Used algorithm is inefficient	Algorithm can be changed to be more efficient and optimal.
Coding level	Loop invariants: Reading end of day currency exchange rate in a million trade record loop to convert price of each trade from Euro to US$.	Onetime currency exchange rate can be read outside the trade record loop and the read value can be used inside the loop, i.e. loop invariant statements can be taken outside loop to avoid repeated execution.
	Resource contention: For example process or threads waiting more than required duration for a shared resource like data base connectivity. Another example is inefficient resource locking and release mechanism among multiple process or threads.	Resource management can be optimized: Resources like database connection or locks can be released immediately after use. Resource pools like connection pool or thread pools can be used to minimize time on resource re recreation every time.
Resource/Infrastructure consumption/utilization level	Un optimal resource utilization: Assume that the system has 2 CPUs or the system is a dual core system and there are 2 stages of data processing in	The strategy can be changed to spawn 2 instances of process A to run in parallel to complete stage 1 first and then to spawn 2 instances of process B to run

	an application and the application spawned 2 processes A and B spawned to run a process in each of the CPU. If the total time is 3 hours and if 1st stage is processed by process A and is finished in first 1 hour, then remaining 2 hours only one among 2 CPUs are being consumed wasting 1 CPU resource for 2 hours.	stage 2, thereby consuming both the CPU resource during the entire duration of execution to maximize resource utilization effectively.
	High memory footprint: Assume that the designed system's memory foot print of each user/session is avoidably high and because of accumulating memory when 50+ users logged in, the application slows down due to very high memory usage.	Memory footprint can be reduced to as minimum as possible for each user/session so that the side effects of high memory consumption slowing down system performance can be minimized.

6. Quantitative estimation and impact study of performance gain from multiple causes and respective solutions

There can be multiple causes and multiple solutions for each cause to the given performance problem. Many applications in general may have performance bottlenecks due to various reasons at various levels and stages and performance gain is possible from respective solutions. Thus the opportunity to make changes and gain performance may be scattered at multiple points from beginning of processing to end of processing throughout the application. For example following changes specific to a specific application may yield performance benefits. Co locating database and application server, few core logic changes, data writes changes to batch writes and data read changes to reduce number of round trips between application server and database server.

Each of the above changes may bring some performance benefits. It is essential to analyze, model, measure or estimate quantitatively the amount of performance gain each change may yield, the tradeoffs with each changes, the cost and effort of each changes, relative benefits, risks of introducing bugs and functional and operational impacts has to be studied. According to the analysis and impact study outcome, implementation of proposed changes can be undertaken. Thus the recommendation should not only contain solutions or tuning parameters or changes and benefits but also should provide adequate support data like quantitative benefits (performance gain) achievable by each change, tradeoffs, effort, cost of making those changes, complexity, risks for each change so that customer get adequate information from the recommendations to do cost benefit analysis and make an informed and calculated decision.

7. Management issues and cost benefit analysis

Most of the time performance issues are identified at much later stage of developments like

7.1. During load and volume testing

7.2. Unexpected peak demand during production

Thus performance engineers are left with limited options of

7.3. Limited timelines to fix the performance issue

7.4. Constrained to make only minor architectural or design changes or deviations as larger changes may require

7.4.1. Longer testing cycles

7.4.2. Higher risk of introducing regression bugs.

7.4.3. Higher effort and hence requires more time for the change

7.4.4. Longer effort and hence higher cost of change

7.4.5. Larger change to the existing system implies low realization of return on investments made on the existing system.

7.5. Engineers can make only limited technical changes

7.5.1. Which requires less effort in terms of time and manpower

7.5.2. Less risk in terms of breaking functionalities or causing new functional bug.

7.5.3. Lower cost by avoiding new third party products, licenses or IPs

8. Case study illustrating approach, drill down to root causes, change parameters and relationship between parameters.

Below sections illustrates with a case study application, how a top down approach described above has identified data access as a cause. How a further breakup of data access in to smaller stages within data access pointed to data transportation and how a further drill down in to data transportation pointed to low pre fetch size (record numbers) as one of root causes for slow performance. How further quantitative analysis could explore a reciprocal or hyperbolic relation between total time to fetch records and pre fetch size and thereby optimum pre fetch size that can be recommended to resolve the performance issue.

8.1. About the case study application

8.1.1. Functional description

The bank acts as a custodian of assets deposited in the bank by their customers or account holders. The case study application facilitates bank's customers to view through web browser across internet and through bank's portal various reports of their assets like positions, balances, corporate actions of companies where bank's customers had invested in, their NAV, interest income etc.

8.1.2. Scenarios or use cases having low performance

The graphical user interface has many links for user to click and view reports with one link for a report. The response time to view majority of the reports in a browser was about 1.5 to 2 minutes and the bank's expectation or requirement is below 10 seconds response time. A response time of less than 3 to 6 seconds will be the expectations as a better user experience.

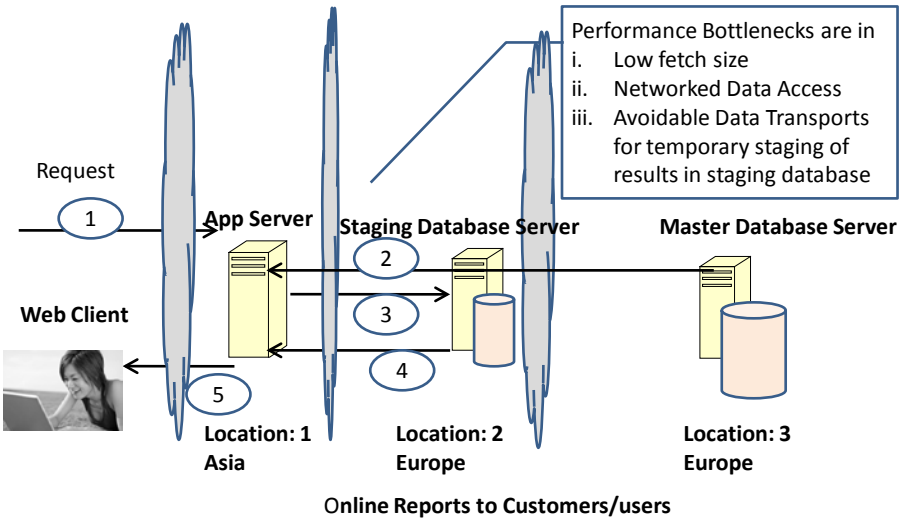
8.1.3. High level view of the system

Architecturally this is a 3 tier web application with the following technology stack.

- JavaScript, JSP, HTML, SmartClient for presentation layer
- Java application deployed in BEA WebLogic application server.
- Data base is Oracle and Oracle global data warehouse.
- Data access mechanism is through Java JDBC APIs

The deployment used and data flow is shown in figure 1 below.

Fig 1: Existing deployment view and data flow



1 to 5 is the data flow sequence after each request

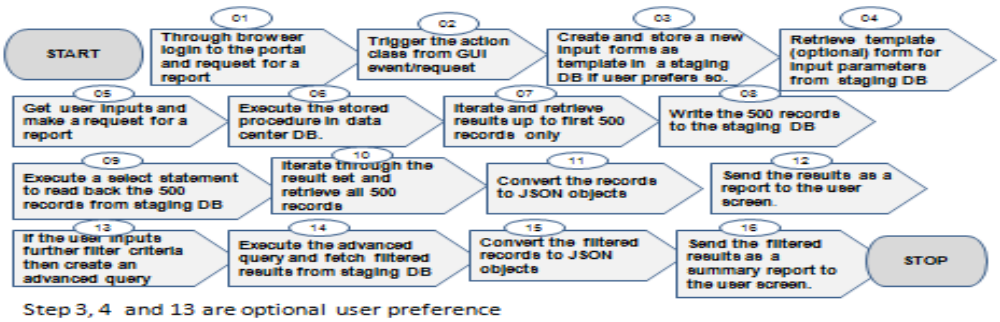
Data flow sequence to serve any user request for reports is depicted in figure 1. Application server fetches data records from global data centre in Europe (location 3) and first 500 records are temporarily parked in a staging data base in another different country in Europe (location 2). Using a generated SQL query as per user requests to further filter the data application server in Asia (location 1) fetches data records and presented as reports to the user browser.

8.2. Breakup of various stages in the end to end process flow in the system

From the time user login and enters requests for some report the flow is divided in to various stages as shown in figure 2. The flow depicted is common across all reports. Some of the use cases are

- User creates and stores an input template form, which can be used to input parameters for the report.
- User either uses a template created earlier or gives input in new form and retrieves a report.

Fig 2: Blocks of processing stages



Step 3, 4 and 13 are optional user preference

Figure depicts overall major processing steps in the case study application and not a flow for a single use case. Depending upon user options and use case only some of the above stages gets executed.

The elapsed time taken for each stage for various reports is measured by code instrumentations and profiling by manual code changes to log time differences across various processing stages. Most of the reports size is about 500 records or limited to 500 records while few summary reports are of size less than10 records. The measured response time to view various reports of size 500+ records ranged from 1 minute to 2 minutes.

8.3. Inference from deployment architecture and time measured across various stages in process flow

From the measured elapsed time for each stage depicted in figure 2 and across various stages in process flow, we could observe that higher

proportion of time was spent on data access from database in data centre and from staging database i.e. stage 6, 7, 9 and 10 in figure 2.

From the deployment architecture in figure 1, we could see that application server where the data was retrieved and processed was located at different geography from where the data base was. Network latency added delay in data access which could be reduced by about one half as measured in the case study by co locating application server and database.

Figure 3 gives further breakup of time between query execution and iteration through result set and retrieval of all records.

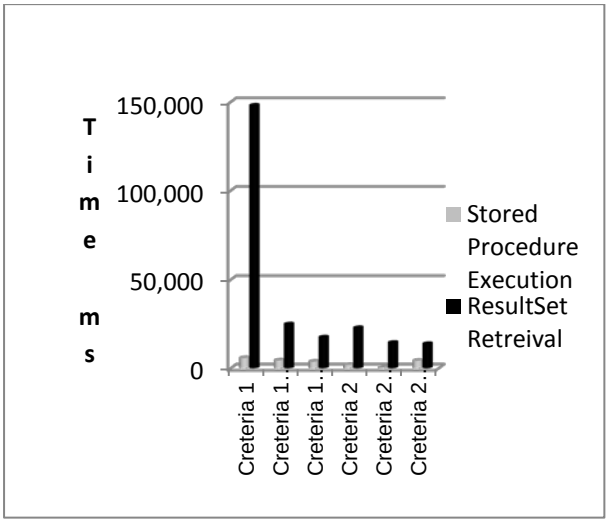


Fig: 3a: Stored procedure execution time and result set retrieval time. Result set has about 502 records.

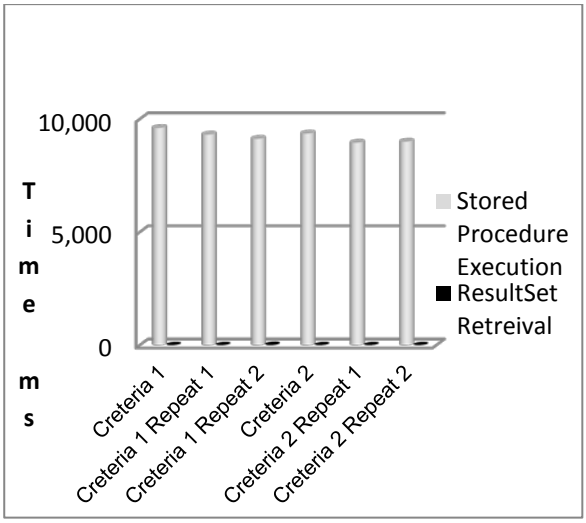


Fig: 3b: Stored procedure execution time and result set retrieval time. Result set has about 5 records.

Figure 3 shows time measured in two stages as

1. Stage 1: Time to execute a callable statement and obtaining a [1]ResultSet, i.e. stored procedure execution time as measured to execute a Java statement resultSet = *cstatement.executeQuery() ;*
2. Stage 2: Time to iterate through the above ResultSet and extract all data records as java objects, i.e. result set retrieval time as measured to iterate through result set and retrieval of all data records as Java objects. i.e. the Java statement

while (resultSet.next()) { //code to extract field objects from result set object }

From the query execution time and results set retrieval time breakups in figure 3 we could see that, when number of records (502) are more (figure 3A), data record retrieval from result sets takes higher proportion of time compared to query execution time. When the number of records (5) is less (figure 3B), query execution time is of higher proportion compared to retrieval time. For the case study application majority of reports are about 500+ records as in figure 3A. Thus at entire application level data retrieval (transport)

consumes higher proportion of time than query execution as well as any other operation or process.

Thus any effort on query optimization and database tuning for this application is not going to give any considerable performance benefit. Hence the focus of optimization in this application should be on optimizing on data transportation and reducing number of round trips between application and the database server rather than on the SQL query optimization.

Having narrowed down to data transport as a cause for delay, let us focus on overall data transportations scenarios in the case study application. From the process breakup (figure 2) and dataflow diagram (figure 1), we could see that for the purpose of doing further advanced filtering of data from already retrieved data records using SQL queries, data are again stored in a staging database. Based on user entered filter criteria a query is formed and filtered data is retrieved from staging database. This staging of data in another temporary database adds additional write to and read from staging data base adding to avoidable delays in data transport. This is a case of logical inefficiency because further advanced filtering can be done without staging the data in staging database.

8.4. Further drill down on data transportation to isolate root cause for higher data transportation time

8.4.1. Pre fetching rows from database to application server

In general a SQL query ResultSet object (a Java JDBC object in application layer) may point to 0 or 1 or thousands or even millions of records which needs to be pulled from database by iterating sequentially through the ResultSet using resultSetObject.next() call. When a SQL query is executed in a database through Java JDBC APIs running inside any application server, almost all database drivers also called as resource adopters have a feature to pre fetch more than just one database record to the ResultSet object of the application even if the request is to iterate and extract one single next record. Subsequent call or execution of the statement “resultSetObject.next()” will extract from pre fetched data records from resultSetObject and not from database server to avoid transportation delay across network between application server and the database server. Oracle has a default pre fetch size of 10 records and Sybase has 20 records.

8.4.2. Configuring pre fetch size

JDBC APIs like setFetchSize(int size) or configurations through application servers are only recommendations to the database driver and the driver may or may not enforce the recommendations. The JDBC APIs to get the pre fetch size i.e. getFetchSize() can give only the recommended value and not the effective value actually used by the driver.

8.4.3. How to identify effective pre fetch size

A simpler and practical way to know what is the effective pre fetch size is to measure the time taken to iterate through and retrieve each record and plot this series of time to fetch consecutive single record from ResultSet object against the sequence number of each consecutive record retrieved and look for peaks in the series. The peak occurs when all the pre fetched and cached records in result set object were already retrieved and to retrieve the requested record, the driver program has to fetch next set of records from database server and not from the local cache in result set object

Figure 4 shows the elapsed time to extract each consecutive record in the case study application, i.e. time to execute the statement **“resultSet.next()”** which is inside a while loop as in the code section below.

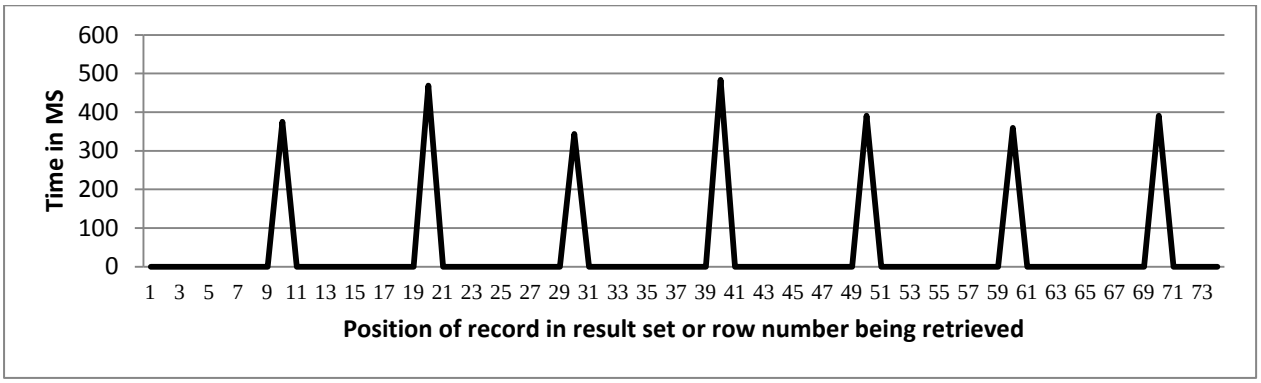
```
resultSet = cstatement.executeQuery() ;  //stored procedure execution time

while( resultSet.next() ) {     // record retrieval time plotted in Y axis and loop index in X axis

  //code to extract java objects from resultSet object

}
```

Fig 4: Retrieval time of consecutive records from result set



From figure 4, we can observe that, time to retrieve each consecutive record was nearly zero milli seconds and negligible, but suddenly increased and peaked to about 300 to 500 ms for every 11th record. This implies that the JDBC driver pre fetched in batches of 10 records at a time and records 1 to 10 are retrieved from local resultSet object which is in memory retrieval and hence nearly zero milli seconds. However when the application attempted to iterate and extract beyond 10th or 20th or 30th etc record, the driver again pre fetched next 10 records from database server which required one more round trip to database server from application server and had to transport a batch of 10 records of data even if the request was for next single record which explained the peak in time of about 300 to 500 ms to fetch 11th, 21st and 31rd records respectively. When we changed the pre fetch size to higher number by configuring pre fetch size in app server or through JDBC APIs, the effective pre fetch size was still 10. This is because setting pre fetch size through JDBC APIs or through app server deployment descriptors are just a recommendation to the driver and not guaranteed and has not changed the effective pre fetch size.

8.4.4. How to set and enforce desired pre fetch size

In the case study application when we used the Oracle JDBC extension class like OracleConnection, OracleResultSet and when we set the fetch size through Oracle extended APIs of these classes we were able to set and achieve desired pre fetch size. The effective pre fetch size can be experimentally observed by similar elapsed time versus row sequence number of records, wherein we could observe the peaks at every 21st record when pre fetch size was set to 20 through Oracle JDBC extension APIs.

8.5. Root cause for low performance in the case study application

In the case study application the query execution time was less compared to records retrieval time. Since effective pre fetch size was the default 10 records and most of the reports were of 500 records, each report needed 50 round trips between database and application server which was identified to be the root cause for low performance.

Thus in the case study application for reports retrieval use cases many reports were of size 500+ records. From the processing stage breakup perspective, among many stages from user request to reports display data access stage was the main performance bottleneck. If we further break up data access in to 2 parts (I) time for query execution and (ii) time to retrieve data records, time to retrieve data records took more time. If we further break up data records retrieval, it was data transportation by multiple round trips between application server and database server which took higher proportion of time. Thus multiple round trips due to low pre fetch size were found to be the root cause for the slow performance in viewing reports.

8.6. Breakup of various stages of data retrieval to application server from database server using SQL query from within a Java JDBC application

Let us try to break up various internal processing steps from the time the application server issues query execute request through JDBC API till the query result set data base records of the executed query from database server are retrieved as

collection of java objects in the application server or client.

From coding perspective using JDBC APIs, above data records retrieval involves two stages

1. Stage 1: Time to execute a statement or prepared statement or callable statement and obtaining a ResultSet, i.e. time to execute a Java statement

   resultSet = *cstatement.executeQuery() ;*

2. Stage 2: Time to iterate record by record through the above ResultSet object which is a pointer or handler to all the records of the query results and extract all data records as java objects, i.e. result set retrieval time from the code section below

   while (resultSet.next()) { //code to extract field objects from result set object }

However internally there may be many underlying processing steps and stages, which are explored below.

Assume that a java application deployed in an application server is making a JDBC call to execute a SQL query which has "N" records in the result set. N may vary from 0 to several millions and if N is very high because of the limitation of memory size in app server, the driver cannot bring and hold all the records in one shot in its result set object but may hold a cursor, a handler in the result set and may fetch a fraction of N records on demand. Thus the driver has to make multiple trips to the database driver when the application iterates through the ResultSet object and retrieves all the N records of the query. Since multiple trips across network is time consuming most of the database drivers optimize by pre fetching a finite number of records ahead even if the application iterates and request for just 1 record. Oracle has a default pre fetch size (f) of 10 records and it may vary with other databases like Sybase, DB2 and MSSQL server.

The breakup of various smaller stages in executing and retrieving "N" records to the application server from database server can be expressed as in Eq. 1 below. If we consider all the select statements to fulfill a use case in an application, then N may include records from multiple select SQL statements; however our focus here is the total read time of N records from a onetime single select statement execution and from single ResultSet Object.

$$T = \sum_{i=1}^{n} (r_i + e_i + a_j + t_i + c_i) \qquad (1)$$

Where

T = Total time to retrieve all N records of the result set to the application server as java objects from the database server.

The subscript "i" can change from 1 to n, where n is the number of round trips the driver had made to bring all the records to the ResultSet object pre fetching 'f' records in each trip. If 'N' is the total number of result set records for the select query and 'f' the effective pre fetch size, then n = N/f + (1 if (N modulus f > 0)). For example if the query result set has N = 502 records and the effective pre fetch size (f=10) then the driver may make 51 (50 to bring 500 and 1 more trip to bring the residual 2 records).

r_i is the elapsed time to make a request from app server to the database server during the i_{th} trip.

e_i is the query execution time spent in the SQL engine running in database server.

e = Hard parse time + soft parse time + search time to select each record meeting the select query criteria + traversal time to locate the searched record + seek time to fetch and join records from multiple tables + time to read records from disk store to the memory of SQL engine.

The SQL engine may do a onetime execution, when i = 1 and may retrieve and cache the records in SQL engine cache also called as database server side caching. Subsequent retrieval of records may be from the SQL engine cache in the database server. This execution time includes hard parse time and a soft parse time. [3]Parse time includes loading SQL statement to memory, a onetime and first time syntax verification of the SQL statement, authorization to access the tables, creating and optimizing execution plan and actual execution time of the select call. Hard parse is relatively more expensive in terms of CPU time. Execution time e will increase if more number of tables are joined, large numbers of records are joined, search on non indexed fields etc.

a_j is the access time spent during the j th iteration by the SQL engine in the database server to retrieve a fraction or all of the N records to SQL engine cache from the database server disk store. SQL engine caching size j referred to as server side cache may be different from the pre fetch size f which is client (app server) side cache. Maximum number of iterations to retrieve all N records from database store to database engine, i.e. upper limit of j depends up on SQL engine's caching capacity in database server whereas the upper limit of i depend upon on the JDBC driver's effective pre fetch size (f) and application server memory availability.

t_i is the elapsed time per trip to transport f records from database server to the application server in i^{th} trip.

$t_i = f1(\beta_i , h_i , b_i, a_i)$
(2)

Where transport time t_i on ith trip is a function f1 of β_i network bandwidth, which will be of the order of few megabytes per second. This transport time decreases with increasing β

h_i number of network hops in ith trip. The application server and database server in data center may sometime be in different location and may even be in different country and hence data has to be transported by more than one network hops. It should be noted that it is not the distance between app server and the database server, but the number of network hops which is relatively significant and important factor in transportation time.

b_i is the number of bytes of data being transported during ith trip. This is the sum of all individual field size in bytes in a record multiplied by the by pre fetch size f (the number of records transported in i^{th} trip.

a_i on wire network bandwidth available due to bandwidth being shared by many different application or process concurrently.

Data transportation time can be expressed as

$t_i = \sum_{h=0}^{h=k} t_i^{k(\beta)}$
(3)

Where, k is the total number of network hops the system has made to transport the pre fetched f records during i th trip between application server and database server. Obviously k = 0, when both application server and database server are same, which is generally not the case in real production level systems. Dependency β in the equation (3) implies that the network bandwidth may vary between each network hop and hence the transportation time may vary in each hop.

c_i is the elapsed time to convert from data base specific data types to java data typed objects. This includes conversion of data in each cell or field of each record and for all the f records, the application fetched in i^{th} trip. For result set in a single select query, since the record structure (number of columns or fields, data type in each column) is same for all the f records, the subscript i in c_i can be removed. In the equation subscript i is retained because the record structure may vary across different select queries in the application. The parameter c may include marshaling data base objects on wire and un-marshaling database specific objects to Java objects by the database driver. Some driver may cause unusually large delays or even error if incompatible data types between database and Java are converted.

In Eq. 1 the factors e_i query execution time and t_i transportation time are two major time consuming factors and in any application either one or both of them may be the dominant time consuming factor. The optimization on either query or transportation or both can be focused accordingly.

Since in our case study application, data transportation time is of higher proportion and found to be the root cause as described in section 8.5, let us focus and elaborate on data transportation components and terms of Eq. 1

8.7. Relation between Elapsed Time and Pre Fetch Size

If we split the data access in application JDBC layer in to two parts as described in section 8.4.4

- stage 1: statement execution time and
- stage 2: data extraction time from query result set which includes predominantly data transportation time to transport data records in batches of f records being pre fetched during each trip and isolate and

measure only the data transportation time, then

$T_2 = \sum_{i=1}^{r} t_i \ \alpha \ R$ (4) and

$T_2 = \sum_{i=1}^{r} t_i \ \alpha \ f$ (5) where LHS is the total transportation time and R is the number of round trips between application server and database to fetch all records of the query result set. In other words total retrieval time is proportional to number of round trips Eq. 4 and also proportional to the size f of data being transported in each trip Eq. 5.

Combining equation (4) & (5), we have

$T_2 = K_1R + K_2f$ (6) Since R = N/f if N is integral multiples of f, otherwise R = N/f + 1 where, N is the total number of records for a given query. Rewriting Eq. 6, we have

$T_2 = K_1N/f + K_2f + K_3 + K_4$ (N modulus f) (7)

Where

K_1 is the proportionality constant of time variation with number of roundtrips when f is the number of records being fetched in each round trip between database and app server. K_1 is average time spent per trip when the data size is of f records.

K_2 is the proportionality constant of time variation with data size. K_2 is the average time to transport data size of f records.

K_3 is the average time per trip, when data size is of (N modulus f) records

K_4 is the average time to transport data of size equivalent to (N modulus f) records

First and third term of Eq. 7 are time spent due to N/f and 1 round trips respectively to transport data records and K_1 and K_3 are proportionality constants of time per trip when size is f and (N modulus f) respectively. Similarly second and fourth terms of Eq. 7 are time spent on transportation due to data size and K_2 and K_4 are proportionality constants of time spent to transport data of size 1 record in single trip. K2 and K4 can be considered as almost equal.

K_1, K_2, K_3 and K_4 can be determined by regression by collecting elapsed time data for various values of number of round trips and data size.

First term in Eq. 7 is rectangular hyperbolic or reciprocal variation of time with pre fetch size, while second and fourth terms are near linear with f. The curve between elapsed time T2 and the pre fetch size f will be either rectangular hyperbolic or linear depending upon which of the two component viz. number of round trips or data size is dominant in Eq. 7. Let us find out the dominant component and term in Eq. 7 from general observations as well as from the case study application measurements in next section.

8.7.1. Comparison of impact of record size and number of round trips on performance

Table 2 gives the data transportation time measured in the case study application for different values of number of round trips and different values of pre fetch size.

Table 2: Impact of data size and number of round trips on data transportation performance

	When pre fetch size f = 10 records	**When pre fetch size f = 252 records**
Impact of data size on performance: Average time per trip to retrieve f records in one trip between app server and database server	450 milli seconds (per 1 trip)	650 milli seconds (per 1 trip)
Impact of number of round trips on performance: Total time to retrieve all 502 records in multiple round trips with pre fetching f records per trip between app server and database server.	14 seconds (in 51 round trips)	2 seconds (in 2 round trips)

From table 2 we could see that, the reduction in time due to reduction in number of round trips between app server and database server is very high (from 14 seconds to 2 seconds) compared to minor increase of time of few milli seconds (from 450 milli seconds to 650 milli second) due to increase in data size. Thus the total time to retrieve all records in a result set is pre dominantly determined by number of round trips expressed in Eq. 4. Since round trips R = N/f where N is the total number of records in a result set and f the pre fetch size after ignoring the data size factor and also the time to transport the residual last 1 trip with N modulus f records. I.e. ignoring all terms in Eq. 7 except the first term. We can write retrieval time T2 as a function f2 of number of round trips neglecting the effect of data size expressed in Eq. 5 and 7, i.e.

$T_2 = f2(R) = f2(N/f)$

$T_2 \alpha N/f$

$T_2 = K_1 N/f$
(8)

$K_1 = f3 (\beta, h)$ (9)

Where K_1 is the proportionality constant of time per round trip when the data size is of f records in each trip. K_1 is assumed a function f3 of network bandwidth β and number of network hops h. Eq. 8 obtained after approximations described above on Eq. 7 is the reciprocal or rectangular hyperbola relation between T2 and f.

8.7.2. Impact of network bandwidth and network hops on performance

Table 3 gives the data transportation time measured between two different network paths from data base to i) app server deployed in Asia and ii) app server deployed in Europe.

Table 3: Impact of network bandwidth and network hops on data transportation performance

Effective Pre fetch size	Time to transport 502 records of query Q with Data centre in Europe and Application Server in Asia	Time to transport 502 records of query Q with Data centre in a country in Europe and Application Server in another country in Europe
10 records	14 seconds	7 seconds
50 records	8 seconds	4 seconds

From table 3 we could see that when all other parameters remains same, time to transport across networks between Europe to Asia is twice that of time to transport records across networks between one European country and another European country. Only difference between these two cases is number of network hops and network bandwidth. We could observe almost a parallel curve of nearly half the time for data transport between the two European countries as compared to transport time between the European country and the Asian country (figure 5).

8.7.3. Rectangular hyperbolic decrease of total elapsed time to transport a set of records with increasing pre-fetch size

Eq. 8 implies a reciprocal or rectangular hyperbolic relation between T_2 total time to transport all (N) records of a result set and the pre fetch size f. We can see from Eq. 8, that when f tends to zero, time tends to infinity and when f tends to infinity time tends to zero. Thus asymptotes are parallel to Y (Time T_2) and X (fetch size f) axis.

Figure 5a is a theoretical curve of Eq. 8 with assumed value of $K_1 = 0.2$. Pre fetch size f is plotted in x-axis and (0.2 * N/f) * 1000 in y-axis. N is taken as 502 and 1000 is multiplied to show time in micro seconds. In the figure 5a only positive values of 'f' is plotted as practically negative f has no meaning.

Figure 5b is the experimental curve measured from the case study application and shows how the total elapsed time "T_2" plotted in vertical y-axis to fetch a report containing about 502 records from database to application server reduces with increasing pre fetch size "f" plotted in horizontal x-axis.

We can qualitatively see the non linear rectangular hyperbolic shape and trend of T_2 vs. f curve measured from the case study application. When the pre fetch size is 0, then the application can

never fetch any data and hence takes infinite amount of time and the curve will be parallel to (T2) y-axis. Similarly when the pre fetch size f is relatively high compared to N, for example 200 in our case study application, application has to make 3 round trips to fetch 502 records and if pre fetch size is increased by 1 unit, the application again needs only 3 round trips to fetch all 502 records. I.e. decrease in number of round trips is 0 per unit increase in pre fetch size when f = 200 or in other words slope is nearly zero and parallel to x- axis at f = 200. In other words the magnitude of slope of the curve is very high or the slope tends to very high negative or minus infinity when f tends to 0 and the magnitude of slope rapidly but smoothly reduces to zero, when f tends to N.

Thus the experimental curve 5b has similar properties of theoretical curve 5a of rectangular hyperbola and the Eq. 8.

Another property of T_2 vs. f reciprocal curve is the decreasing slope with increasing f. The slope important and worth to study because

1. Negative slope indicates that elapsed time T_2 to fetch records decreases with increasing f.
2. Decreasing magnitude of slope with increasing f indicates that the rate at which elapsed time T_2 decreases or the gain in performance for a unit increase in fetch size f is relatively high when f is small and the gain in performance for a unit increase in f is relatively small when f is large.

We can see how the slope decreases with increasing f by comparing the slopes at different points on the curve in figure 5b which is detailed from case study measurements in the next section.

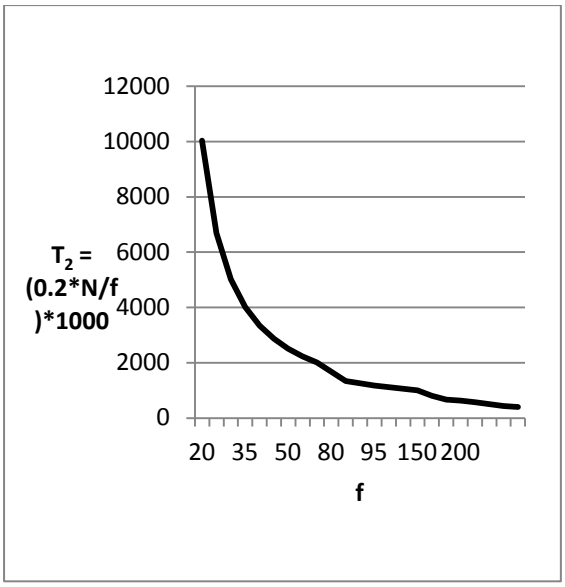


*Fig 5a: Theoretical curve of T_2 = (0.2*N/f)*1000 Vs f*

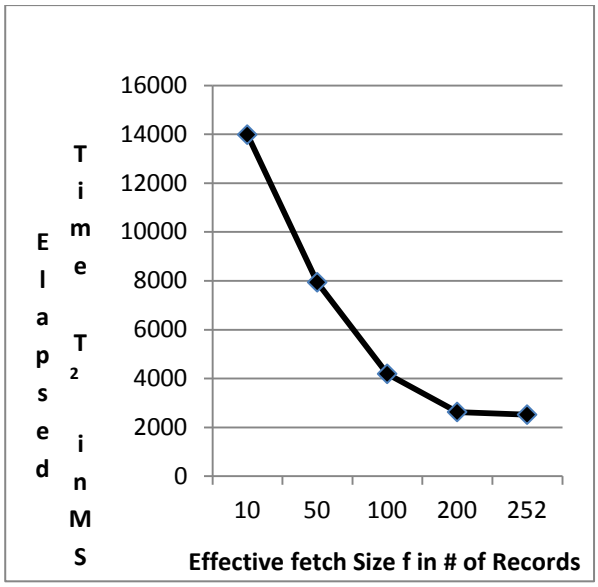


Fig 5b: Experimental curve from case study showing decrease in elapsed time (T_2) with effective fetch size (f)

8.7.4. Understanding the reason for decreasing magnitude of slope with increasing pre fetch size of transportation time vs pre fetch size curve

Decreasing slope with increasing f in T_2 vs. f curve (fig 5b) for a given N can be understood by calculating and focusing on the decrease in number of round trips achievable per unit decrease in f at different points in the curve for increasing values of f from low value of f to high value of f. From simple illustrative calculation shown in table 4 for a given N = 502, we can see that when f is low there is a high reduction in number of round trips and hence higher gain in performance even for a small or unit increase in f but when f is higher less reduction in number of round trips and hence less significant performance gain by same quantitative or unit increase in f.

Table 4: Illustration of decreasing slope with increasing pre fetch size

Seq #	f_1- Effective Pre-fetch size in number of records	n_1 -Number of round trips to fetch 502 records when pre fetch size is f_1	f_2- Effective pre fetch size increased by 1	n_2 -Number of round trips to fetch 502 records when pre fetch size is f_2	(n_1-n_2) - Decrease in number of round trips to fetch 502 records per unit increase in f at f_1.	Slope of T_2vs f curve. Decrease in time in ms per increase in f by 1 record. Assuming average of 400ms per round trip.
1	0	Infinity	1	502	Infinity	Infinity
2	1	502	2	251	251	100400
3	10	51	11	46	5	2000
4	100	6	101	5	1	400
5	168	3	169	3	0	0
6	200	3	201	3	0	0

Column 6 in table 4 shows how the reduction in number of round trips per unit increase in f at various values of f to fetch 502 records decreases with increasing f from infinity to zero rapidly as f is increased from zero to 168. One can see from table 4 that for a total number of records of 502, when f is low like 1 record, even for a small increase of f to 2 records, the reduction in number of round trips to fetch all 502 records reduces from 502 to 251. One can compare this with reduction in number of roundtrips by only 5 for a same unit increase in f, when f is 10 and a reduction of just 1 round trip by unit increase in f when f is 100. Thus the performance gain per unit increase in pre fetch size is very high when f is low and the performance gain reduces rapidly as f increases. Performance gain is very low per unit increase in f when f is higher and reaches zero after f >= 168 for N = 502.

To generalize for any values of total number of records to retrieve N, one can say that, when the ratio (N/f) between number of records N and pre fetch size f is high, even a small change in f will bring large benefits in performance. When the ratio N/f is small, even large change in f will not get considerable performance benefit. This is in consistent with the reciprocal nature of the Eq. (8) and the theoretical curve of figure 5a.

Figure 6 shows the plot of the slope dT_2/df in y-axis and f in x-axis, where T_2 is the total elapsed time to transport about 502 records observed in the case study application and f is the effective pre fetch size in number of records.

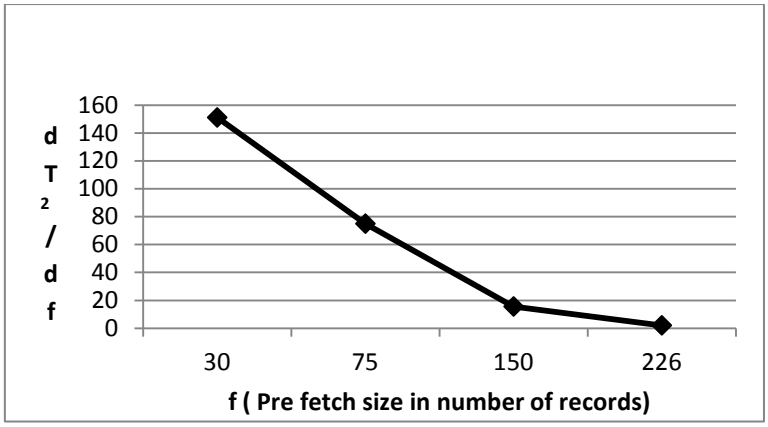


Figure 6: dT_2/df (Slope of Elapsed Time With Pre Fetch Size)

Figure 6 is the derivative of curve of figure 5b, where we can observe that the rate of decrease of elapsed time with pre fetch size (slope) decreases with increasing pre fetch size.

8.7.5. Differences between theoretical rectangular hyperbolic curve and actual transportation time vs. fetch size curve

Though there are many similarities between theoretical rectangular hyperbolic curve as in Eq. or figure 5a and actual T_2 Vs. f curve, there are many differences. Few differences are listed below.

Eq. 8 is only an approximation after removing size factors and other terms in Eq. 7, hence actual T_2 vs f curve will have slight lift towards higher T_2 which increases with f at each point in the curve when compared with theoretical curve (ref fig. 5a and fig. 5b). The lift is due to size factor of slight increase in time to transport higher data size due to higher number of records (f).

For theoretical rectangular hyperbolic curve, the 'f' in Eq. 8 has to be continuous, whereas practically f is discrete.

Also at certain values of f for a given N, T_2 may not decrease even when f is increased unless increase in f results in decrease in number of round trips. For example assume that N = 502, and f is increased from 251 to 252. Number of round trips is same and equal to 2 for both values of f. Comparing time T_2 when f = 251 and f = 252, one can see that, time to transport 251 records twice may be of same value or may not decrease when compared with time to transport 252 records in first trip and remaining 250 records in 2nd trip.

Thus the reciprocal or rectangular hyperbolic nature of T_2 vs. f is only a dominant trend and an approximation and not an absolute relationship between T_2 and f. However rectangular hyperbolic trend between transportation time and pre fetch size can be treated as a generic dominant and approximate trend for any application accessing relational data through JDBC and will help in performance tuning and deciding optimal pre fetch size. If N is known and K1 is determined, then this relation can give an approximate estimate of quantifiable expected performance gain achievable for various values of f without much trial and error.

8.8. Trade off between higher pre fetch size and memory consumption

Though performance gain can be achieved by reducing the number of round trips between application server and database servers by increasing the pre fetch size, higher pre fetch size requires higher memory allocation in application server. In the case study application, when the pre fetch size was increased from default 10 records to 500 records as most of the report size was about 500 records, the application server was found to crash with out of memory error exception. Thus there is a tradeoff between performance gain by higher pre fetch size and higher memory consumption.

8.9. Threshold pre fetch size

From the figure 5 and 6 as well as from the rectangular hyperbolic nature of T_2 Vs f curve, we

can see that, the curve (very high negative slope) is almost parallel to Y axis (T_2 axis) when X (f) is near zero. Thus there is a rapid decrease in elapsed time (T_2) with increase in pre fetch size (f) when pre fetch size was low. However beyond certain f, the curve (has very low or near zero slope) is almost parallel to X axis (f axis) and hence very less or near zero decrease in elapsed time (T_2) per unit increase in f when f is large. For the case study application we can see that from figure 5b, the curve became parallel to x axis (f axis) for values of f above 200 and from figure 6 we can see that the slope (rate of decrease of elapsed time with pre fetch size) already became zero when f = 226. Thus there going to be no significant performance gain by increasing pre fetch beyond 226 records. Since pre fetch size of 226 requires 3 round trips (226 + 226 + 50) to fetch all 502 records, with 502/3 = 167.3, i.e 168 (168 + 168 + 166) fetch size, we can fetch all 502 records in same 3 round trips. The performance loss due to slight increase in time required to fetch when the data size increases due to higher fetch size being less significant compared to performance loss due to increased number of round trips as round trip is being considered as a main tuning parameter. With same number of round trip lesser pre fetch size will have lesser memory footprint in app server hence 168 should be the preferred pre fetch size than 226. Pre fetch size of 168 can be considered as the threshold or optimal pre fetch size for this case study application where the number of required records N is limited to 502.

9. Recommendations for tuning the case study application

Based on elicitation and extraction of application architecture, deployment architecture, data and process flow logic and analysis of the same, time measurements by code instrumentation, profiling of various major stages of execution, identification of causes and solutions and relative benefits of solutions the following recommendations were made.

Table 5: Proposed performance tuning recommendations for the case study application

Identified performance bottleneck	**Suggested solution as changes**	**Level of abstraction of the change and estimated effort in person days**	**Trade off or factors to consider before making implementation decision**	**Estimated and measured benefits in trial implementation of recommended solution/changes.**
Application layer is making avoidable multiple (50) round trips between app server and data base as the pre fetch size was default 10.	Increase the pre fetch size to the threshold value of 168. Use Oracle Connection, Oracle Statement and Oracle Result Set to effectively set the pre fetch size as setting pre fetch size in application server or setting pre fetch size through JDBC Connection, Statement and Result set were ineffective.	Code level. 2	Higher pre fetch size demands higher heap memory in application. Using Oracle specific extension APIs Vs generic APIs.	Measured benefit from average 70 seconds to 25 seconds per report retrieval.
Application server and data center are at geographically different locations leading to higher network latency to access data due to more number of network hops and	Move application server to the location of data center to co-locate data and application consuming the data. Moving database to application location is expensive and hence the	Deployment Architecture. 3	Moving application server from existing Asia location to Europe location where data center is implies relocating application IT team to Europe or	From 25 seconds to 12 seconds per report retrieval as time was found to reduce by about one half for retrieving many reports.

hence delays.	suggestion of moving application server to data center.		doing remote support.	
There was an extra write and read operation due to staging result sets in a staging database	Change the application logic to directly read the data from main data base and avoid staging database.	Logic and Design. 5	With retrieved data temporarily staged in database further filtering on retrieved records can be easily done by SQL query filters. Avoiding staging database implies doing data filter in Java.	This extra write and read operation was estimated to consume 4 seconds per report retrieval. Estimated time gain is from 12 seconds to 8 seconds.
User views only first 500 records, whereas the stored procedure extracts all and more than 500 records meeting the query criteria. Some queries had even 40 thousand records.	Change the stored procedure to limit the number of records to first 500 among all records meeting the query criteria.	Code/SQL script: 3	Unused records are being fetched in database wasting time and memory. This change has no trade off.	Not estimated.
Field (column) size of record in the table is not limited to required size like 50 characters length in Java. Instead it uses table field default size of 4000 characters per VarChar2 field. This causes huge memory allocation when records are fetched into app server from database.	In Table definition use varchar2(50) instead of varchar2. This indicates to the driver to allocate only 50 characters instead of 4000 characters per column of type Varchar2.	Data Model/Table Structure. 2	More than necessary field width is consuming avoidable memory.	This reduces the memory demand in app server while reserving memory to store result set records and enable us to use higher pre fetch size.

10. Conclusion

Re engineering an existing application in production is different from engineering for developing a new application for performance or for any other quality of service. There is a growing need to re engineer many business applications already deployed and serving in production for higher performance due to increasing demands. Performance problems can be in any of several processing or data flow stages like user inputs, network, processing, data access etc. Solutions can be applied as changes at architecture, design, framework, logic and coding level. An inquisitive and experimental approach starting with higher abstraction (e.g. architecture) level view to identify causes and solutions, then a drilled down (e.g. design) level causes and solution and further drill down to identify root causes and respective solutions will help. More than one factor can cause performance degradations hence estimating relative benefits will be helpful. Cost benefit analysis requires measured or sampled or estimated quantitative benefits and trade offs for each of the solutions. It is common for multi tired business applications with data access as the common performance bottlenecks area. Though there can be a slight increase in transportation time with data size, there is a dominant reciprocal or rectangular hyperbolic relationship between total transportation time to retrieve result set records of a query and the pre fetch size, the number of records the database driver brings to the client and caches at client side from database server. Data records retrieval time reduces more rapidly with increasing pre fetch size when the ratio of number of records to retrieve to

pre fetch size is high however the performance gain is negligible when this ratio is low. There is a threshold value of pre fetch size, where transportation time Vs pre fetch size curve appears to approach a point of inflexion where slope tends to zero when pre fetch size is increased further. Beyond this point, increasing pre fetch size will not bring any considerable performance benefit.

11. Acknowledgments

Author thanks Anupam Sengupta of Tata Consultancy Services Ltd for his effort in reviewing this paper and his valuable and pointed technical comments has helped me to make many corrections and improvements from initial version. Author thanks Banisha M, Syed Irfan Pasha and Saiaparna Kunala of HCL technologies who were part of case study application development team and helped me in analyzing the case study application. Banisha explained the implemented architecture and design of the case study application. Syed developed modules to measure time to iterate through and retrieve each record from ResultSet to identify the effective pre fetch size and Saiaparna instrumented the application to measure breakups between query execution time, query results retrieval time and to profile time across various stages of processing which helped us to isolate the performance hot spots.

References

[1] Oracle Corporation, Java™ Platform, Enterprise Edition 6. API Specification, US, 2009-11. Ref: http://docs.oracle.com/javaee/6/api/

[2] Oracle® Database JDBC Java API Reference 11g Release 2. US, 2009. Ref: http://docs.oracle.com/cd/E18283_01/appdev.112/e13995/oracle/jdbc/OracleResultSet.html

[3] Donald K. Burleson, Oracle Tuning The Definitive Reference Second Edition, US, Oracle In-Focus series #32, 2010. Ref: http://www.dba-oracle.com/t_hard_vs_soft_parse_parsing.htm

Modeling a Distributed Database System for Voters Registration in Nigeria

Olabode Olatubosun

Business Information System
University of Botswana

Abstract.
The Independent National Electoral Commission, Nigeria is characterized for managing large volume of dispersed data making distributed data processing a necessity. When voter rolls are error-ridden and a quarter of eligible voters cannot vote, registration laws are not only failing their primary function of ensuring that voters are qualified to vote but also acting as barriers to citizens democratic participation. The traditional voter registration methods employed by many developing countries for periodic elections have many associated problems such as incomplete, inconsistent, unavailability and erroneous records. This article presents an application of distributed database system for a complete and continuing voter registration in Nigeria. The system has its component parts physically stored in a number of distinct real databases at a number of distinct sites. Each site has its own local real databases, its own local users, its own local DBMS and transaction management software including its own local locking, logging, recovery, replication, fragmentation, e.t.c. software and its own local data communication manager. Distributing data across sites within state and local government allow voters data to be resident where they are generated or most needed, but still accessible from other sites within the state and local government areas. Java and Oracle were the developmental platform of the system. Some important relations for the systems were presented and possible management transaction and operation models were presented. The system require a Unix/windows NT operation system in a network environment such as provided by communication networks in Nigeria and an internet connection.

Keywords: Distributed system, Fragmentation, Replication , Smart Card, Voter Registration

1.0 INTRODUCTION

In a well-functioning democracy, voting should be protected as a fundamental citizenship right and responsibility [1]. Accurately registering every eligible voter to vote is a necessary step toward protecting this right, yet a very high percentage of eligible Nigerian citizens voters are not registered and many people are registered inaccurately or engage in multiple registration.

An "automated" voter registration system is one in which government offices, including social service offices, collect and transfer voter registrations to election officials without using separate paper forms but direct capture machines [2]. These offices enter registration data into their computers and transfer them electronically, in a format that election officials can securely review and upload directly into their voter registration database systems. Many developed States such as Arkansas, California, Georgia, Kentucky, Michigan, New Jersey, North Carolina, South Carolina, South Dakota, Texas and even Nigeria are already at this stage. An alternative approach which is an "online" registration system is one that allows individuals to submit a voter registration application over the Internet. Six states such as Arizona, Colorado, Kansas, Louisiana, Oregon, and Washington currently have online systems in place for individuals who have a driver's license or non-driver's identification card. At least five more states like California, Indiana, Nevada, North Carolina, and Utah are developing similar systems.

The Federal Republic of Nigeria, with an area of 923,769 square kilometers (made up of 909,890 square kilometers of land area and 13,879 square kilometers of water area. The 2006 national population census puts the country population at 140,431,790 people. The country is subdivided in 39 states plus Federal Capital Territory (Abuja). The states are further divided into a total of 774 local government areas [3]. Democracy in Nigeria is still in the struggle to leapfrog to an ideal democratic setting. Usually, every election in Nigeria is associated with fraud of various dimensions. Nigeria is in the 4th republic yet it is embarking on new voters register. One thing is for sure, there are no effective registration system to enhance complete, accurate consistent and continuous voters register which is a pre-requisite for a credible election.

As [1] notes, when voter rolls are error-ridden and a large number of eligible voters cannot vote, registration laws are not only failing their primary function of ensuring that voters are qualified to vote but also acting as barriers to citizens democratic participation. In Nigeria, many eligible voters are disfranchise and a left with no other choices.

To address these deficiencies with regards to voters registration, one needs a robust, efficient and effective information technology system, that can systematically register every eligible voter in Nigeria and give them information about voting mechanics and electoral choices. Such a modern and universal voter registration approach would include the design of a Distributed Database System for the Independent National Electoral Commission (INEC) that is saddled with the responsibility of ensuring a complete, accurate and effective voter registration list of citizen and participations. The framework for the database architecture and structure presented. This article shall also present models for the transaction, operational and processes inherent in the system.

The distribution of data in a network or decentralized computer system offers several attractive advantages over the centralization of data at a single computer. These advantages include increased data reliability; faster, localized access to data; and the potential for upward scaling of data capacity [4].

2.0 Literature Review

In [5] and [6] a distributed database system consist of a collection of sites, connected together via some kind of communications network, in which, each site is a full database system site in its own right, but the sites have agreed to work together so that a user at any site can access data anywhere in the network exactly as if the data were all stored at the user's own site. This follows that a distributed database is really a kind of virtual database, whose component parts are physically stored in a number of distinct real databases at a number of distinct sites and each site has its own local real databases, its own local users, its own local DBMS and transaction management software including its own local locking, logging, recovery e.t.c. software and its own local data communication manager. In particular, a given user can perform operations on data at that user's own local site exactly as if that site did not participate in the distributed system at all.

The Distributing data across sites within state and local government will allows those data to reside where they are generated or most needed, but still to be accessible from other sites in the state and local government areas. Keeping multiple copies of the database across different sites will allows continuous database operations even when one site is affected by a natural disaster, such as flood, fire, or earthquake or manmade incidences. Distributed database systems is structured geographically or administratively distributed data spread across multiple database systems. [7] opined that the central function of a distributed database system is to provide access to data while maintaining the integrity and consistency of that data. The system must have the ability to support large numbers of users without sacrificing performance. Higher reliability and availability in the presence of equipment and network failures are requirements for mission critical enterprise data systems. These

requirements are often at odds with each other, leading to solutions that compromise between availability, consistency, scalability and performance.

Many researchers who have used distributed database system for the management of their enterprise data, including [8]. A computer-based healthcare record system being Developed for Boston's Health care for the Homeless Program (BHCHP) uses client-server and Distributed database technologies to enhance the delivery of healthcare to patients of this unusual population. The needs of physicians, nurses and social workers are specifically addressed in the application interface so that an integrated approach to healthcare for this population can be facilitated. Usually, patients and their providers have unique medical information needs that are supported by both database and applications technology. To integrate the information capabilities with the actual practice of providers of care to the homeless, the computer-based record system was designed for remote and portable use over regular phone lines. An initial standalone system was used at one Major BHCHP site of care. The project describes methods for creating a secure, accessible, and scalable computer-based medical record using client-server, distributed database design.

Also in the study of [9], Studies of voter turnout across states find that those with more facilitative registration laws have higher turnout rates. Eliminating registration barriers altogether is estimated to raise voter participation rates by up to 10%. The article presents panel estimates of the effects of introducing registration that exploits changes in registration laws and turnout within states. New York and Ohio imposed registration requirements on all of their counties in 1965 and 1977, respectively. Also in the study they find out that the introduction of registration to counties that did not previously require registration decreased participation over the long term by three to five percentage points. Though significant, this is lower than estimates of the effects of registration from cross-sectional studies and suggests that expectations about the effects of registration reforms on turnout may be overstated.

In the article of [10] State Congress enacted the National Voter Registration Act (NVRA) of 1993 in order to establish procedures that will increase the number of eligible citizens who register to vote in elections for Federal office. The NVRA mandates simultaneous voter registration and registration updates with driver's license applications and renewals; use of mail registration forms; the establishment of agency-based registration forms at state offices, including public assistance and unemployment compensation offices; and restrictions on purging of voter registration rolls and States without voter registration requirements (ND) and states which permit election-day registration at the polling place are exempted from the requirements of NVRA .

2.1 Nigerian Voters Registration.

The Federal Republic of Nigeria in 2010 had electronic voter registration. Each registration centre was equipped with a laptop computer, fingerprint scanner, a camera for photo passport and a printer for quality voters card. When a person comes in to apply for registration, his/her biometrics details are captured digitally, that is, digital images of his/her photograph, fingerprints and signature will be taken/captured using the Data Capture Machine (DCM). The intention of government is to have a clean, complete, permanent, and updated list of voters through the adoption of biometrics technology in the registration process. Any Nigerian citizen who is at least eighteen (18) years of age. A resident of the Nigeria for at least one (1) year and in the place wherein he proposes to vote for at least six (6) months on or before the day of the election; and Not otherwise disqualified by law are eligible to register. Eligible voters personally appear before the registration officer, state his/her name and exact address, specifying the house number, name of street, and local government area. AT the end of the voter registration exercise, capture data where processed and released. How accurate and authentic this

records are leaves more questions to be answered.

The chairman of INEC in one of his speeches claimed that the 2011 national voters registration will be in line with that of 2008 voters registration in Bangladesh. Bangladesh employed 30,000 direct capture machine and took between 8 to 11 months to embark on the exercise. The Bangladesh Army has selected MegaMatcher SDK multi-biometric technology to identify duplicate registrations in the nation's voter database. The Bangladesh Voter Registration Project registered more than 80 million citizens using biometric face and fingerprint technology. After evaluating a number of biometric technologies for their duplicate search system, the Bangladesh Army determined that MegaMatcher from Neurotechnology was able to identify more duplicate registrations with a higher degree of accuracy than any other system tested [11]. According to the article, System integrator Dohatec New Media was hired to help design and implement the MegaMatcher-based system. To date, more than 48 million voter registration records have been matched. The Dohatec Biometrics Fusion Server system uses MegaMatcher Client to generate templates from face and fingerprint images that were captured with a BIO-Key system, then the match technology is used to search the database and identify duplicate records. Bangladesh runs MegaMatcher on Microsoft Windows XP and Microsoft Windows Server with Microsoft SQL Server as the back-end database. MegaMatcher provides the high speed and reliability required for the development of national-scale automated fingerprint identification systems (AFIS) and multi-biometric face/fingerprint identification systems. Suitable for both civil and forensic use, the system includes both fingerprint and face identification engines with a fusion algorithm that allows the two technologies to work together to provide very fast 1:N matching with even higher reliability than AFIS or facial recognition alone.

2.2 Theoretical Background of Distributed Database.

The data which frequently resides on multiple sites inside an organization might be managed by several Database Management Systems for multiple reasons such as scalability, performance, access and management, [12, 5, 13] Thus, the information requirements for executing transactions and answering questions might not reside in a single site. Distributed Database Management Systems deal with distributed database as a single logical database, and the principles and techniques of Database Management Systems are still applicable to the distributed one; although the distributed one has special characteristics. A distributed database management system is a software that support the transparent creation, access and manipulation of interrelated data located at the different sites of a computer network [14]. Furthermore, [15] describe a distributed Database Management System (DDBMS) governs the storage and storage of logically related data over interconnected computer system in which both data and processing are distributed among several sites. Each site of the network has autonomous processing capability and can perform local applications. Each site also has the potential to participate in the execution of global application, which is to improve the accessibility, compatibility and performance of a distributed database while preserving the appearance of a centralized database management system [14]. Moreover, Distributed database system are very complex systems that have many interrelated objectives of transparency, heterogeneity, autonomy, high degree of function, extensibility and openness and optimized performance. It should be noted however, that, data allocation is done largely at the discretion of the database designer or database administrator [5, 14, 16].

A typical DDBMS consist of four major component [16], of the Local DBMS component responsible for controlling the local data at each site that has a database and has its own local system catalog that store information about the data held at that site. It contains the Data Communication (DC) component which is a software that enables all sites to communicate

with each other and the Global System Catalog (GSC) with functionality to hold information specific to the distributed nature of the system such as fragmentation and allocation schemas [16, 15] and the Distributed DBMS component is the controlling unit of the entire system. A distributed system requires functional characteristics that can be grouped and described as transparency features. These were discussed in [15] as distributed, transaction, failure, performance and heterogeneity transparency.

The database is physically distributed across the data sites by fragmenting and replicating the data [17]. Given a relational database schema, fragmentation subdivides each relation into horizontal or vertical partitions. Horizontal fragmentation of a relation is accomplished by a selection operation which places each tuple of the relation in a different partition based on a fragmentation predicate. Vertical fragmentation, Divides a relation into a number of fragments by projecting over its attributes. Fragmentation is desirable because it enables the placement of data in close proximity to its place of use, thus potentially reducing transmission cost, and it reduces the size of relations that are involved in user queries. Based on the user access patterns, each of the fragments may also be replicated. This is preferable when the same data are accessed from applications that run at a number of sites. In this case, it may be more cost-effective to duplicate the data at a number of sites rather than continuously moving it between them [6].

[4] consider a network of interconnected computers. Each computer, known as a node in the network, contains a distributed database management system (DDBMS) and a possibly redundant portion of the database. Data are logically viewed in the relational data model. The unit of data distribution is a relation. The DDBMS will maintain system directories so that each query will receive a nonredundant consistent mapping of its required data. Data transmission in the network is via communication links. The data transmission cost between any two nodes is defined as a linear function $(X) = c_0 + c_1X$, where X is the amount of data transmitted. The our cost measure in units of time. The constant c_0 represents an initial start-up time for each separate transmission

A relation r is fragmented into fragments $r_1, r_2, \dots, r_n$ either horizontally or vertically. According to [12, 5, 13] horizontal fragmentation involves a relation r is divided into a number of subsets, $r_1, r_2, \dots, r_n$. Each tuple of relation r must belong to at least one of the fragments, so that the original relation can be reconstructed. Canonically, a horizontal fragment can be defined as a selection operation on the global relation r. That is, a predicate p_i to construct fragment r_i.

$$r_i = \sigma p_i(r)$$

and to reconstruct the relation r, the union of all the fragment is taken, thus

$$r = r_1 \cup r_2 \cup \dots \cup r_n$$

In turn, vertical fragmentation of $r(R)$ involves the definition of several subsets of attributes $R_1, R_2, \dots, R_n$ of the schema R so that

$$R = R_1 \cup R_2 \cup \dots \cup R_n$$

each fragment r_I of r is defined then by

$$r_i = \prod_{k=1}^{n} R_i(r)$$

And to reconstruct r, the natural join is taken as

$$r = r_1 \bowtie r_2 \bowtie r_3 \dots \bowtie r_n$$

One way to ensure a successful relation reconstruction is to include the primary-key attributes of R in each R_i.

[16] included the mixed fragmentation of a relation consisting of a horizontal fragment that is subsequently vertically fragmented or a vertically fragmented that is then horizontally fragmented. This approach is defined using the selection and projection operations of relational algebra. Given a relation R, a mixed fragment is defined as

$$\sigma_p(\pi_{a1,\ \dots,an}(R))$$
$$\pi_p\left(\sigma_{a1,\ \dots,an}(R)\right)$$

[5] presented this scenario in form fundamental fragmentation rules as:
Rule 1: Completeness. If a relation instance R is decomposed into fragments $R1, R2, \dots Rn$, each

data item that can be found in R must appear in at least one fragment. This rule is necessary to ensure that there is no loss of data during fragmentation.
Rule 2: Reconstruction. It must be possible to define a relational operation that will reconstruct the relation R from the fragments. This rule ensure that functional dependencies are preserved.
Rule 3: Disjointness. If data item d_i appears in fragment R_i, then it should not appear in any other fragment. Vertical fragmentation is the exception to this rule, where primary key attributes must be repeated to allow reconstruction. This rule ensures minimal data redundancy.

Also in [12] access to various data item in a distributed system is usually accomplished through transaction, which must preserve the ACID properties [16]. The transaction can either be local or global transaction.

2.4 Voter Registration Model

Voter Registration is a procedure required of prospective voters and used to establish their identity and place of residence prior to an election so that they are certified as eligible to vote in a precinct. The purpose of voters registration is to dimish opportunities for election day vote fraud, [20]. Voter registration exists for the fundamental reasons of Registration information used to control who votes. Only those who are eligible to vote can register, and that eligibility is verified when the individual registers to vote. Also, registration information is used to authenticate voters when they participate at poll sites. Thus, voter registration exists to control access, and to prevent voter fraud. Other reason is that Registration information is used for election management and for other election administration tasks. Voter registration lists contain the addresses of those eligible and registered, and that information is used for many purposes ranging from provision of polling places to insuring that every voter receives the ballot they are supposed to receive when they go to vote. Voter registration is also used to maintain historical information to manage voter lists going forward and to provide evidentiary information in case of a challenge to the outcome of an election [21].

Today, voter registration is a massive, complex, and dynamic database problem. At the national level, one must keep track of something more than 70 million registered voters and in a database with more than 70 million records, and many pieces of information about each registered voter, typographical and other errors are inevitable. Furthermore, Nigerian population is dynamic; voters move very frequently each year, according to data from the 2006 National Population Census, new voters are constantly entering the picture, by becoming eligible to vote i.e. turning 18 and also voters are constantly leaving the eligible electorate, either by death or other reasons. There are some specific proposed policies to achieve goals of complete voters registration: vis:

a. uniform age of sixteen for advance voter registration
b. registration of high school students during compulsory government examinations
c. automatic registration of citizens obtaining driver's licenses and learner's permits, marriages permit and any other forms of government registration
d. print and online voter guides; and
e. television and radio time devoted to election information.

In the case study, Nigeria, Periodic registers in which a register could be established for a single electoral event or for any electoral events occurring within a defined period is very common. Periodic registers generally require voters to register a new and previous registrations are not taken into account. Although the use of modern technology, was involved in the exercise, data capture, storage, retrieval, update, dissemination of information is still a challenge. Voters information are characterized with inconsistency, duplication, redundancy and also the integrity of data and efficiency is very poor. Identity documents containing photographs, signatures or

finger/thumb prints are usually generated using specialised systems designed to produce identity cards while the subject is present. In these cases, textual information is printed on hardcopy using data either provided on the spot or data extracted from a database. The voter usually signs this hardcopy record, and/or makes a fingerprint or thumbprint. The operator places the hardcopy printout, including the signature and/or finger/thumb print in the device, and takes a photograph of the person. The device then prints an identity card including a copy of the printed data, the signature and/or finger/thumb print and the photograph. The card is usually laminated and the integrity may be improved by including tamper-evident security devices such as holograms or embedded print to make it difficult to forge or alter the card. This phenomenon may be continuous register that can be constantly kept up-to-date by amending and adding voter records whenever necessary. A database systems can be used to easily update records and add new records, as well as keep track of amended and deleted records.

Voter register databases system can be used to manage subject data. When photographs, signatures or finger/thumb prints have been digitised and stored in a database, various methods exist to manipulate such data types and also have it printed on identity documents by the voter register database system. In recent times, identity cards can be produced in the form of Smart cards, incorporating magnetic strips or data chips to store electronic data about the person who is the subject of the card. This data may include bio-identification data. the smart cards can be used with smart card readers and bio-identification readers such as finger print scanners to automatically verify a person's identity. Smart cards can be “read only” cards that simply contain information about the subject or can be “read-write” cards, which have the information contained on the card updated as the cards are used. For example, a read-write card used to verify a person's right to vote could, once used, be recorded as having been used for that election, so that it could not be used for voting in that election again. Where smart cards are used in polling places, they could be used to replace current methods of recording that a person has voted. Where a voter uses a smart card at a polling place to verify his or her right to vote, the smart card reader could at the same time record that that person had voted and transmit that data to a central database during or after polling. Though provision of smart cards to voters and smart card readers to polling places is expensive, users need to weight the advantages against the expense. Moreover, Smart cards incorporating an electronic identity could also be used for voter registration or voting by computer over the Internet or at a computer kiosk, provided the computer was equipped with a smart card reader. Embedded modules can be used to perform a range of tasks that can assist in reducing instances of fraudulent registration or voting and to identify and delete instances of duplicated voter registration records. Voters Registration in Bangladesh is a good example to emulate.

Software can perform various comparison routines to determine whether a person applying for registration is already registered, perhaps at another address. Electronic searches can be programmed into voter registration databases to identify whether a person applying for registration is already on the register.

In general, the Caltech/MIT Voting Technology Project (VTP) has outlined five basic standards that a voter registration system must meet, Registration information must be accurate and complete, must be immune from fraud, be dynamic and up-to-date, be usable by election officials at polling places and must be easy for eligible individuals to register to vote. Current and future voter registration systems should be assessed relative to these standards. The INEC can be seen as enterprise that is distributed already, at least logically into National, State and local governments and perhaps wards from which it follows that data are distributed already as well because it is expected that each unit of the INEC will naturally maintain data that is relevant to its own operation. The total information asset of the INEC then is thus

splintered into what are sometimes called island of information. The distributed system provides the necessary bridges to connect those islands together

Relational Model:

A formal way of presenting a relation schema. Let $R(f_1: D_1, \ldots, f_n: D_n)$ be a relation schema and for each $f_i, 1 \leq i \leq n$, let Dom_i be the set of values associated with the domain named D_i. An instance of R that satisfies the domain constraints in the schema is a set of tuples with n fields:

$$\{\langle f_1: d_1, \ldots, f_n: d_n\rangle | d_1 \in Dom_1, \ldots, d_n \in Dom_n\}$$

The 5 different relations are required for this system, includes

$\boldsymbol{State}[statecode, statename]$
$\boldsymbol{Lga}[lgacode, lganame, statecode]$
$\boldsymbol{Ward}[wardcode, wardname, lgacode]$
$\boldsymbol{Unit}[unitcode, unitname, Street, wardcode]$
$\boldsymbol{Voter}[regno, fname, lname, othername, sex, datebirth, address, hometown, occupation, employer,$ passport, thumb, $lga, state, unitcode]$

Figure 2 presents the schematic structure of the 5 relations

Notations

Number of States $= i$
Number of Local government area $= j$
Number of wards per local government $= k$
Number of Units per ward in a local government area which may vary $= l$
Therefore we can represent the state, lga , wards and units as:
States $= S_i, \quad i = 1, 2, 3, \ldots, 36$
Local government areas $= L_{ij} \quad i = 1, 2, 3, \ldots, 36; j = 1, 2, \ldots, n$
Wards $= W_{ijk} \quad i = 1, 2, 3, \ldots, 36; j = 1, 2, \ldots, n; k = 1, 2, \ldots, m$
Units $= U^l_{ijk} \quad i = 1, 2, 3, \ldots, 36; j = 1, 2, \ldots, n; k = 1, 2, \ldots, m; l = 1, 2, \ldots, o$

Estimates

Total number of registered voter in state i, lga j, wards k and units Poll Units l is given as $cU^l_{ijk} = count_{l=1:o}(U^l_{ijk})$
Total number of registered voter in state i, lga j, in wards k is given as $cW_{ijk} = \sum_{l=1}^{o} cU^l_{ijk}$
Total number of registered voter in state i,in lga j, is given as $cL_{ij} = \sum_{k=1}^{m} cW_{ijk}$
Total number of registered voter in state i is given as $cS_i = \sum_{j=1}^{n} cL_{ij}$

Sample Queries

Case1- The Database support location transparency

```
SELECT *
FROM Okitpupa
WHERE lgacode = '16-OKP'
UNION
SELECT *
FROM Irele
WHERE lgacode = '16-IRE'
UNION
SELECT *
FROM Akure_South
WHERE lgacode = '16-AKS'
```

Case2- The Database support location transparency

```
SELECT *
FROM KTP NODE W1
WHERE AGE > = 18
UNION
SELECT *
FROM KTP NODE W2
WHERE AGE > = 18
UNION
SELECT *
FROM KTP NODE W3
WHERE AGE > = 18
```

2.5 The Intranet

Intranet standard for exchanging e-mail and publishing web pages are becoming interestingly popular for business use within closed networks called Intranets. A typical intranet is connected to the wider public internet through a firewall with restriction imposed on the types of information that can pass into and out of the intranet [16].

Table 4: Poll Details list

pollunit	pollname	street	ward

Table 1: Voter Register list

votid	pollunit	employer	fname	lname	Othere	dateobirth	originn	birthplace	occupation	passport	thumb
OND14-U01-102	OND14-	Self	Labake	Jayeoba	Joy	12/4/1967	Ondo	Ilutitun	Farmer	OND14-U01-102P	OND14-U01-102T
OND1 U01-103			nle	Adebowale		3/5/1970	Kwara	Mase	Teacher	OND14-U01-103P	OND14-U01-103T
OND1 U01-104			le	Adebowale	Unice	12/12/1976	Ondo	Ilutitun	Teacher	OND14-U01-104P	OND14-U01-104T
OND1 U02-105			l	Adebobaje		5/3/1966	Ondo	Ilutitun	Nurse	OND14-U02-105P	OND14-U02-105T
OND1 U02-106			bora	Orimogunje	Edward	4/2/1978	Ondo	Erinje	Tailor	OND14-U02-106P	OND14-U02-106T
OND1			uyi	Ibikunle		11/13/1975	Ondo	Igbotako	Farmer	OND14-	OND14-

Table 2: State list

StateCode	StateName
ABI	Abia
ADA	Adamawa
AKW	Akwa Ibom
ANA	Anambra
BAU	Bauchi
BAY	Bayelsa
BEN	Benue
BOR	Borno
CRO	Cross River
DEL	Delta
EBO	Ebonyi
EDO	Edo

School, Igbotako

	D14-7	Abodi, Ikoya	Liliken	OND14-W13
	OND14-U08	Jemiken, Ilutitun		
	OND14-U09	Court Hal Erinje		

Table 3: Local Government Areas list

LgaCode	LgaName	StateCode
OND01	Akoko North East	OND
OND02	Akoko North West	OND
003	Akoko South East	OND
004	Akoko south West	OND
005	Akure North	OND
006	Akure South	OND
007	Ese-Odo	OND
008	Idanre	OND
009	Ifedore	OND
010	Ilaje	OND
011	Ile-Oluji-Okeigbo	OND
012	Irele	OND

Table 4: Ward list

WardCode	WardName	LgaCode
OND14-W01	Ilutitun Ward 1	OND14
OND14-W02	Ilutitun Ward II	OND14
OND14-W03	Ilutitun Ward III	OND14
OND14-W04	Igbotako Ward I	OND14
OND14-W05	Igbotako Ward II	OND14
OND14-W06	Igbinsin	OND14
OND14-W07	Okitipupa Ward I	OND14
OND14-W08	Okitpipupa Ward II	OND14
OND14-W09	Irinje Ward I	OND14
OND14-W10	Irinje Ward II	OND14
OND14-W11	Okunmu Ward I	OND14

Figure 2. Database Relational Structure of the voters Registration System

Three tier model which solves the problem of enterprise scalability is proposed with the following layers of architecture.

a. The user interface layer which runs on the end-user's computer (the client)
b. The business logic and data processing layer. This middle tier runs on a server and is often called the application server
c. The DDBMS which stores the data required by the middle tier. This tier may run on a separate server called the database server.

The implementation language is java. Java is a proprietary language developed by Sun Microsystem and currently marketed by Javasoft. According to [19]. The importance of Java language and its related technologies has been increasing for the last few years. Java [22] is a type-safe, object oriented programming language that is interesting because of its potential for building web application (applets) and server application (servlets).

Java as explicitly defined is a simple, object-oriented, distributed, interpreted, robust secure, architecture neural, portable, high-performance, multi-threaded and dynamic language [23].

Conclusion

Voter registration is one of the stages at which there are ample opportunities to manipulate election results. For this reason special efforts should be made to ensure that the voters list is accurate and reliable in other words all eligible voters are listed only once, and eligible. There has been a growing consensus among election officials, scholars, and voting rights advocates that voter registration can be automated to take advantage of new information technologies, making the process more cost-effective, accurate, and efficient for government and voters. I have presented this article in the effort to sensitize stake holders of concerned organization on the need to decentralize voter registration and to make it continuous exercise. Nigeria is a very large country with fairly large population. If many of the eligible voter are not registered, they are automatically disfranchised. Credible voters register is a pre-requisite for a credible election and credible election to a large extent will guarantee good governance which is what has eluded the Nigerian government over the past decade. The recent development in the world of information technology has brought great change in the dynamic world. Information can be processes accurately, transmitted from any place to anywhere via the networks, data can be sparsely processed, managed and secured, etc. Here, we have proposed a distributed database model for continuous voters registration in Nigeria. The cases where distributed database system is implemented is presented, the state of current voter system in Nigeria is also presented and theoretical background of Distributed database system presented. A model for future continuous voter registration in Nigeria is proposed and the transaction and algebraic operation on the databases presented. Java has been suggested as the ideal language for the implementation of the system. If the management and control of voters register can be sparsely managed with embedded forensic application software in a network environment. Then the nation can have a reliable update of voters register.

References

1. R. Robert. "Seeking 100 Percent Voter Registration and Effective Civic Education". Published online in Wiley InterScience www.interscience.wiley.com. National Civic Review 2007. • DOI: 10.1002/ncr.186.
2. P. Christopher. "Voter Registration in a Rigital Rge". Brennan Center for Justice at New York University School of Law. Edited by Wendy Weiser. 2010.
3. Annual Abstract of Statistics. Available at www.nigerianstat.gov.ng. 2009.
4. R. Alan Hevner and S. Bing Yao Query. "Processing In Database Systems". IEEE Transactions On Software Engineering, Vol. Se-5, No. 3. 1979.
5. C.J. Date. An Introduction to Database System, 8th Eds, Pearson Education, Inc. 2004.
6. M.T. Ozsu and P. Valduriez. "Distributed and Parallel Database Systems". In Handbook of Computer Science and

Engineering, A. Tucker (ed.), CRC Press, pages 1093. 1997
7. E. Todd, M. Charles, R. Peter and J. P. Gerald. "The Bengal Database Replication System, Distributed and Parallel Databases", Kluwer Academic Publishers, 9, 187–210. 2001.
8. C. C. Henry, and C. O Barenett,. "Client-server, Distributed Database Strategies in a Health-care Recoed System for a Homeless Population. Journal of the American Medical Informatics Association. Vol. 1, No 2. 1994.
9. A. Stephen and M.K David. "The Introduction of Voter Registration and Its Effect on Turnout". Political Analysis, Vol. 14 Issue: Number 1 p83-100, 18p; 2006. AN 8133809.
10. K. Stephen and J. White. "Did states' motor voter programs help the Democrats?". World Bank. http://mpra.ub.uni-muenchen.de/28052/ MPRA Paper No. 28052, posted 10. January 2011 / 23:56. 1998.
11. Digital Communities. "Bangladesh Voter Registration Project using Biometrics to Detect and Prevent Duplicate Registrations". December 11, 2008 By News Report. http://www.digitalcommunities.com/articles/E-Vote-Bangladesh-Biometric-Voter-Identification-Project.html?utm_source=related&utm_mediumdirect&utm_campaign=EVote-Bangladesh - Biometric-Voter-Identification-Project.
12. M.T. Ozsu and P. Valduriez. Distributed and Parallel Database Systems, In Handbook of Computer Science and Engineering, A. Tucker (ed.), CRC Press, 1997, pages 1093.
13. R. Ramakrishnan and J. Gehrke. "Database Management Systems", McGraw-Hill, 3rd Edition. 2003.
14. L. Sam, T. Teorey and T. Nadeu. Physical Database System, Morgan Kaufmann Publisher. 2007.
15. C. Coronel, M. Steven and R. Peter. Database Principles Fundamental of design, Implementation and Management, 9th Edition, Course Technology. 2011.
16. T. Connolly, B Carolyn and S. Anne. Database Systems: A practical Approach to Design, Implementation, and Management, Third Ed. Addison Wesley. 2010.
17. S. Ceri, B. Pernici and G. Wiederhold. "Distributed Database Design Methodologies". Proceedings of the IEE, pp 533-546. 1987.
19. C. Egyhazy and K. Triantist,. "A Query Processing Algorithm for Distributed Relational Database Systems". The Computer Journal, 31(1). 34-40. 1986.
20 R. Michael Alvarez. Voter Registration: Past, Present and Future. Written Testimony Prepared for the Commission on Federal Election Reform, Caltech/MIT Voting Technology Project. 2005.
21. G. H. Utter, Strickland, R. Ann. "Campaign and Election Reform". A Reference Handbook Contemporary World Issues Publication: Santa Barbara, Calif. ABC-CLIO, 1997.
22. J. B Gosling, G. Steele and G. Bracha. The Java Language Specification. 3rd edition. Addison Wesley, NY. ISBN 0-321-24678-0. 1996.
23. Sun.. Sun Microsystems 1997. Java Home Page http://java.sun.com. 1997. Access May 5, 2011.

A Comparison Between Data Mining Prediction Algorithms for Fault Detection (Case study: Ahanpishegan co.)

Golriz Amooee[1*], Behrouz Minaei-Bidgoli[2], Malihe Bagheri-Dehnavi[3]

[1] Department of Information Technology, University of Qom
P.O. Box 3719676333, No.52, 24th avenue, 30 metri Keyvanfar, Qom, Iran

***Corresponding author**

[2] Department of Computer Engineering, Iran University of Science and Technology
Tehran, Iran

[3] Department of Information Technology, University of Qom
Qom, Iran

Abstract

In the current competitive world, industrial companies seek to manufacture products of higher quality which can be achieved by increasing reliability, maintainability and thus the availability of products. On the other hand, improvement in products lifecycle is necessary for achieving high reliability. Typically, maintenance activities are aimed to reduce failures of industrial machinery and minimize the consequences of such failures. So the industrial companies try to improve their efficiency by using different fault detection techniques. One strategy is to process and analyze previous generated data to predict future failures. The purpose of this paper is to detect wasted parts using different data mining algorithms and compare the accuracy of these algorithms. A combination of thermal and physical characteristics has been used and the algorithms were implemented on Ahanpishegan's current data to estimate the availability of its produced parts.

Keywords: *Data Mining, Fault Detection, Availability, Prediction Algorithms.*

1. Introduction

In today's competitive world, improving reliability, maintainability and thus availability of industrial products becomes a challenging task for many companies. Reports indicated that performance and availability largely depends on reliability and maintainability. [14] There are many solutions to improve the maintenance of complex systems. One solution is corrective actions, including the required repair and maintenance activities after the occurrence of failures and downtimes. Because of the high costs, risks and time consumptions, this method is not appropriate. Another solution is to use systematic and planned repair methods. Although this method prevents from serious failures, but it could be very expensive. [13]

Given the issues above, to reduce costs and increase availability more effectively, it is better to predict errors before occurrence using data analysis. In industrial companies, since the generated data volume is growing there is a major need to process data in real time. New technologies, in term of both software and hardware, provide data collection from different resources, even when production rate is really high. In the company discussed in this paper, several sensors were installed on the devices to collect and review information during the operation. The information includes temperature, pressure and speed. Also a data base management system (DBMS)

is used to control and managed data stored in the data base. [1]

Current methods of data analysis which work based on reviewing the appearance or annual statistical graphs, has many limitations to predict the performance and availability of the produced parts. So today, analysts are attending to use superior patterns to increase availability. In recent years, data mining is considered as one of the common methods for processing and discovering the hidden patterns. Tools such as data warehouses, data mining, and etc, provides new field for production and industry. So that by using these tools, companies can achieve competitive advantages. Specifically, through data mining - extracting hidden information from large data bases - organizations are able to predict future behaviors, and can make decisions based on knowledge. [2] Using data mining techniques to detect errors and inefficiencies, largely increase the capacity of equipments productivity. [5]

In this paper, we used data mining tools to identify defective parts in Ahanpishegan manufactory. Such analysis results in providing high quality products, improving produced parts and thus increase availability. The outline of the paper is as follows:

First we described required concepts briefly. Then we present an overview on the previous researches conducted to identify defective parts. We also frame the steps needed to create a model and apply data mining algorithms on Ahanpishegan's data. Finally, we provide some suggestions to improve the model for further studies.

2. Required concepts

2.1. Data mining

Data mining discovers hidden relationships in data, in fact it is part of a wider process called "knowledge discovery". Knowledge discovery describes the phases which should be done to ensure reaching meaningful results. Using data mining tools does not completely eliminate the need for knowing business, understanding the data, or familiarity with statistical methods. It also does not include clear patterns of knowledge. [2]
Data mining activities are divided into three categories:

1. **Discovery:** Includes the process of searching the database to find hidden patterns without a default preset.

2. **Predictive Modeling:** Includes the process of discovering patterns in databases and use them to predict the future.

3. **Forensic Analysis:** Includes the process of applying extracted patterns to fined unusual elements.

2.2. Prediction Algorithms

The purpose of a prediction algorithm is to forecast future values based on our present records. [3] Some common tools for prediction include: neural networks, regression, Support Vector Machine (SVM), and discriminant analysis. Recently, data mining techniques such as neural networks, fuzzy logic systems, genetic algorithms and rough set theory are used to predict control and failure detection tasks. [5] In this paper, the algorithms will forecast a probability for the given data situation. If the probability is equal to 1 it means the data (part) is normal, otherwise if the probability is equal to 0 the data (part) is considered non-conventional.

2.3. Failure

Failure means having faults, interruption or stop in a system which is shown as a deviation in one or more variable. The most common way to detect, predict and avoid failures is to collect and analyze the information produced during the time of operation and maintenance. [1] This detection, prediction, and avoidance of failures in early stages will increase the availability.

2.4. Fault Detection

Fault detection is defined as detecting abnormal process behaviors. [10] Fault detection techniques are divided into two categories: Model-based approaches [11] and Data-based approaches. [12] Since the construction of comparative models for real-time industrial processes is difficult, therefore model-based fault detection techniques are not as popular as data-based fault detection techniques. In recent years statistical tools such as PCA (Principal

Components Analysis), KPCA (Kernel Principal Component Analysis) and DPCA (Dynamic Principal Components Analysis) are widely developed.

3. Related works

Only a few articles exist in the field of identifying defective parts with the help of data mining tools. In an article Dr Shabestari review various types of defects in casting aluminum parts in the Aluminum Research Center of Iran. He concluded that a proper understanding of the characteristics of defective parts is a necessity for suppliers. He also concludes that the best way to solve this problem is to make a board of defects along with an example of defective parts and label each component with the name of corresponding fault. Mr. Ghandehari et al. used steel grain size to detect defective parts in terms of mechanical properties. They used non-destructive method of abysmal flows Instead of destructive methods of metallographic which was time consuming and costly.

Mr. Alzghoul et al. [1] used Data Stream Management Systems (DSMS) and data stream mining to analyze industrial data with the aim of improving product availability. They used three classification algorithms to investigate the performance and products availability of each algorithm. Algorithms which were used are: Grid-based Classifier, Polygon-based Classifier and One Class Support Vector Machine (OC-SVM). They concluded that OC-SVM accuracy (98%) is better than the two other algorithms.

In addition to this paper, other researches also used data stream mining for machine monitoring and reliability analysis [7], online failure prediction [8] and tool condition monitoring [9].

In other researches, Sankar et al. [10] used non-linear distance metric measured for OC-SVM to detect failures, Or Rabatel *et al.* [13] review monitoring sensors' data to find abnormalities and increase maintainability. Mr. Huang et al. [14] used statistical T2 and PCA methods to detect different failures in thermal power plants.

4. Research method

In this section we describe the approach which was used to predict defective parts. This article aims to identify defective parts by using multiple prediction algorithms and help the firm to maximize its productivity and increase its reliability and availability. Different steps of our work is described in next sections.

4.1. Understanding the data

The research presented in this paper is carried out in collaboration with Ahanpishegan Co., a manufacturer in automotive industry and a producer of car aluminum parts for companies like Sapco, Part tire, etc. This factory produced many parts such as engine bracket, feed bracket, Rear and front side brackets, etc. Due to the wide range of parts, we only focus on engine brackets. The general shape of this part and its relevant sizes, from two different angles, are shown in figure 1 and 2.

4.2. Data Preprocessing

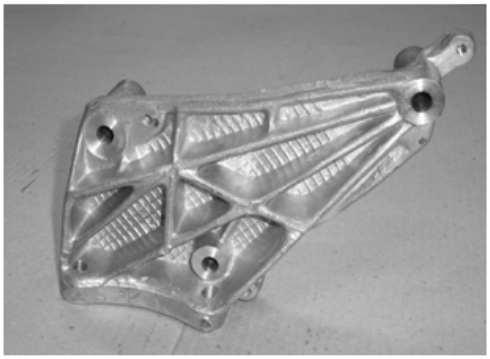
Preparation and data preprocessing are the most important and time consuming parts of data mining. In this step, the data must be converted to the acceptable format of each prediction algorithm. First we find remarkable points about features and proportion of defective part, through interviews with managers and employees. For example, rising temperature has a large impact on corruption, or the rate of defective parts in last months of summer is higher. Then we identify outliers, cleanse data and ignore the constant variables by analyzing the current and past records and multiple interviews with experts and staff. Then we specify important fields which should be used in our prediction algorithms (our next step) to identify defective parts. Selected field for prediction algorithms, outliers and null values are shown in figure 3.

4.3. Creating artificial abnormal data

At this point we need to enter some distorted and corrupted data to measure the influential variables on performance. Also entering this data has a significant role in comparing the accuracy of our prediction algorithms. In this article we entered 10% defective part (100 records of our 1000 records).

4.4. Applying various prediction algorithms

Different kinds of trees such as CHAID, C&R, and QUEST along with other prediction algorithms including neural networks, Bayesian, logistic regression, and SVM has been applied on our data. The results are presented in next session.



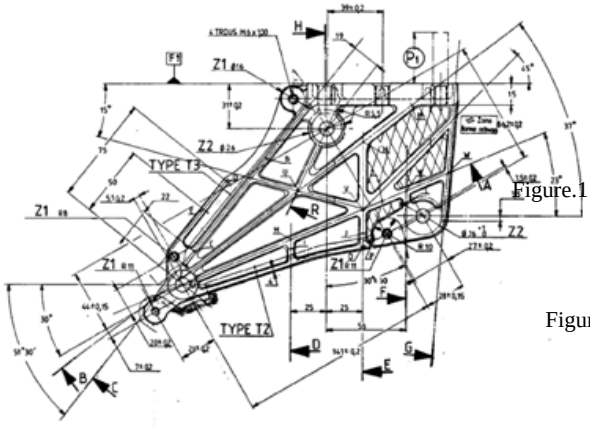
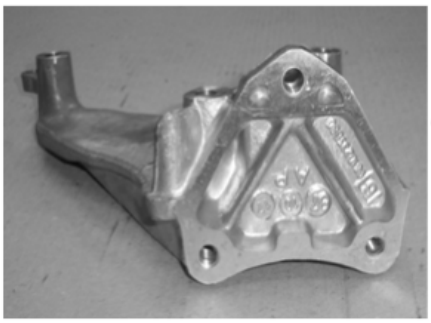
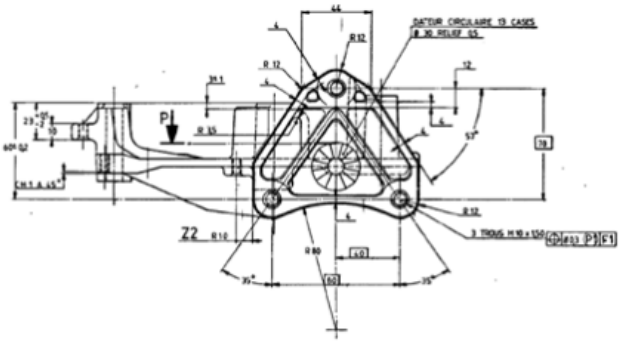


Figure.1 Engine bracket and its sizes f

Figure.2 Engine bracket and its siz





	Field	Graph	Type	Min	Max	Outliers	Null values
1	Mold temprature		range	54.000	5343.000	11	4
2	Melting temprature		range	65.000	2000.000	7	7
3	Hardness		range	5.000	700.000	6	5
4	Machining		range	0.010	756.000	1	9
5	Prevet the black pieces		set	0.000	1.000	--	--
6	Distance between sensitive points and umbilical		range	2.000	200.000	4	12
7	Preventing damage		range	0.000	1.000	5	2
8	Efficiency		set	0.000	1.000	--	--

Figure.3 Final variables for pre

4.5. Investigating each algorithm's accuracy

The results of applying different algorithms are illustrated in figure 4. These algorithms are compared based on their accuracy and processing time. Due to data cleaning and removing noises and outliers, algorithms' accuracies are high.

Based on the result you can see that, SVM has the best processing time and also great overall accuracy. On the other hand algorithms which use trees to create their model needs more time and are sensitive to binary fields, but as you can see C&R and QUEST achieve the highest classification accuracy. A schema of a created C5 tree is shown in appendix A. Neural network is less accurate. Since our fields are numerical neural network has its own difficulties.

4.6. Rule generation using C5

After running different prediction algorithms and evaluating each of their accuracies, at this stage we aim to connect engine bracket features with known failures. So we run a C5 model on our data to generate prediction rules. An example of generated rules is as follows:

1) Mold temperature =< 325.500 and hardness <= 82 and distance between sensitive point and umbilical =< 23.95 then the part is normal.
2) Mold temperature > 325.500 and hardness > 82 and distance between sensitive point and umbilical =< 23.200 then the part is defective.

5. Conclusions and recommendations for future work

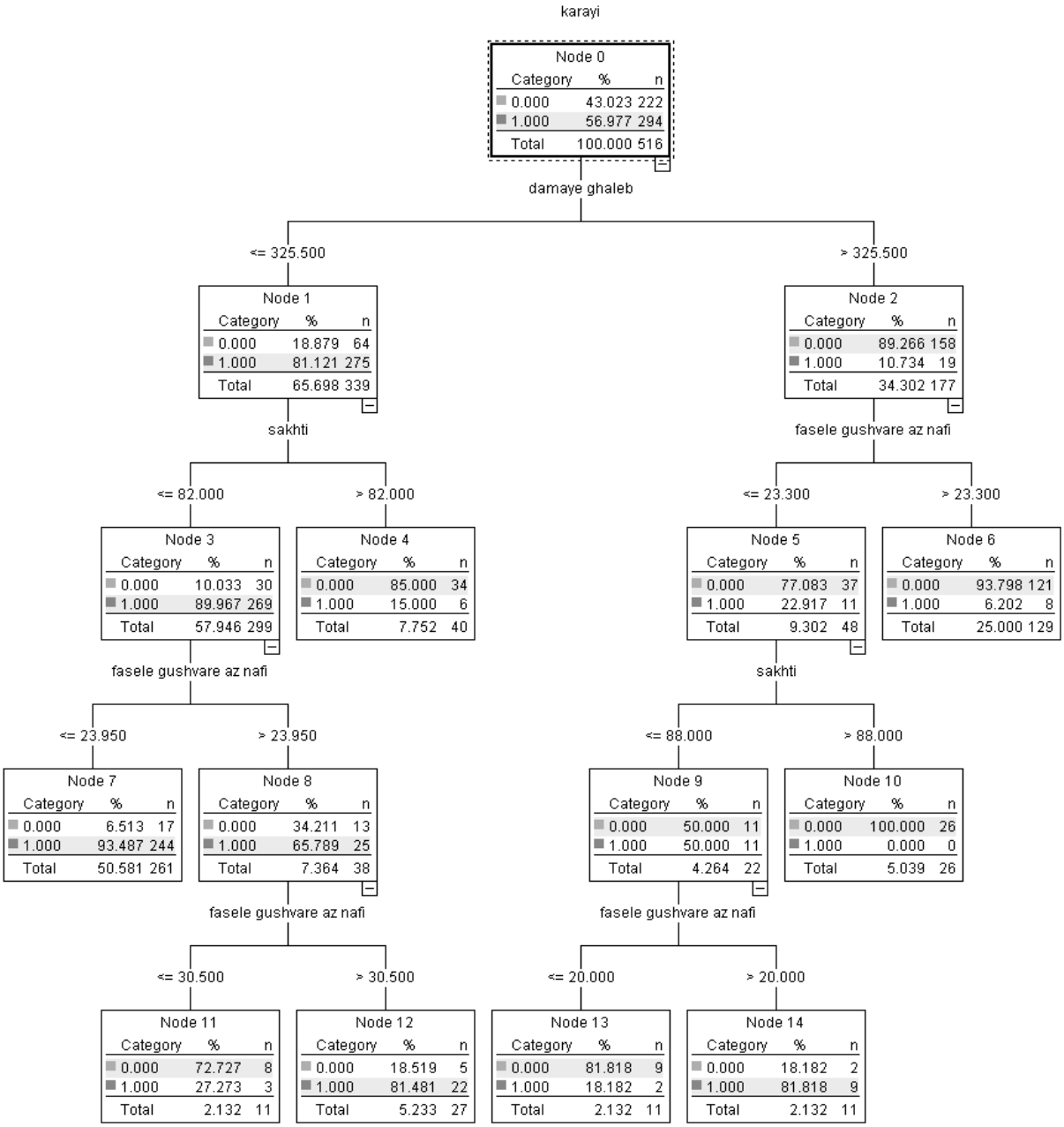
The purpose of this paper is to use data mining tools for identifying defective parts. The research presented in this paper has been carried out in collaboration with Ahanpishegan Co. First we find remarkable points about features and proportion of defective part, through interviews with managers and employees. Then by using an integrated database, identifying outliers, cleaning the data and ignoring the constant variables, we apply different prediction algorithms and compare the results. This paper aims to improve industrial product's reliability, maintainability and thus availability.

We recommend using data stream mining in future works to achieve quicker and more accurate results. Such researches can be performed in other companies such as food industry to measure products quality. Also future works can focus on frequency of failures, the cost of each failure and aim to minimize the consequence of such failures.

Figure.4 A comparison between algorithms

	Algorithm	Processing time (minute)	Accuracy (%)	Unused fields	Area Under Curve
1	CHAID	< 1	88	3	0.94
2	Neural net	< 2	79	7	0.95
5	C&R Tree	<5	92	6	0.92
6	QUEST	<4	92	6	0.92
7	Bayesian Network	<3	89	7	0.94
8	Logistic regression	< 1	89	7	0.93
9	SVM	< 1	90	7	0.92

Appendix A


karayi
Node 0
Category % n
0.000 43.023 222
1.000 56.977 294
Total 100.000 516
damaye ghaleb
<= 325.500
> 325.500
Node 1
Category % n
0.000 18.879 64
1.000 81.121 275
Total 65.698 339
sakhti
Node 2
Category % n
0.000 89.266 158
1.000 10.734 19
Total 34.302 177
fasele gushvare az nafi
<= 82.000
> 82.000
<= 23.300
> 23.300
Node 3
Category % n
0.000 10.033 30
1.000 89.967 269
Total 57.946 299
Node 4
Category % n
0.000 85.000 34
1.000 15.000 6
Total 7.752 40
Node 5
Category % n
0.000 77.083 37
1.000 22.917 11
Total 9.302 48
Node 6
Category % n
0.000 93.798 121
1.000 6.202 8
Total 25.000 129
fasele gushvare az nafi
sakhti
<= 23.950
> 23.950
<= 88.000
> 88.000
Node 7
Category % n
0.000 6.513 17
1.000 93.487 244
Total 50.581 261
Node 8
Category % n
0.000 34.211 13
1.000 65.789 25
Total 7.364 38
Node 9
Category % n
0.000 50.000 11
1.000 50.000 11
Total 4.264 22
Node 10
Category % n
0.000 100.000 26
1.000 0.000 0
Total 5.039 26
fasele gushvare az nafi
fasele gushvare az nafi
<= 30.500
> 30.500
<= 20.000
> 20.000
Node 11
Category % n
0.000 72.727 8
1.000 27.273 3
Total 2.132 11
Node 12
Category % n
0.000 18.519 5
1.000 81.481 22
Total 5.233 27
Node 13
Category % n
0.000 81.818 9
1.000 18.182 2
Total 2.132 11
Node 14
Category % n
0.000 18.182 2
1.000 81.818 9
Total 2.132 11

Acknowledgments

The authors would like to thank the referees for their valuable comments and suggestions, which greatly enhanced the clarity of this paper. We also would like to thank all the people that contributed with their expertise and work to the realisation of this research.

References

[1] Alzghoul, A., Löfstrand, M.," Increasing availability of industrial systems through data stream mining", Computers & Industrial Engineering (2010), doi: 10.1016/j.cie.2010.10.008.

[2] Chris Rygielski, Jyun-Cheng Wang , David C. Yen." Data mining techniques for customer Relationship management" Technology in Society 24 (2002) 483–502.

[3] E.W.T. Ngai, Li Xiu , D.C.K. Chau." Application of data mining techniques in customer relationship management: A literature review and classification" Expert Systems with Applications 36 (2009) 2592-2602

[4] P. Yang, S.S. Liu, Fault Diagnosis for boilers in thermal power plant by data mining,in: Proceedings of Eighth International Conference on Control, Automation,Robotics and Vision, Kunming, China, December 6–9, 2004.

[5] Kai-Ying Chen , Long-Sheng Chen, Mu-Chen Chen , Chia-Lung Lee." Using SVM based method for equipment fault detection in a thermal power plant" Computers in Industry 62 (2011) 42–50.

[6] Karacal, S.C., *Data stream mining for machine reliability* in *IIE Annual Conference and Exhibition.* 2006.

[7] Karacal, S.C. *Mining machine data streams using statistical process monitoring techniques.* in *IIE Annual Conference and Expo*. 2007.

[8] Youree, R., et al. *A multivariate statistical analysis technique for on-line fault prediction.* 2008.

[9] Karacal, C., S. Cho, and W. Yu, *Sensor stream mining for tool condition monitoring.* Computers and Industrial Engineering, 2009: p. 1429-1433.

[10] Sankar Mahadevan, Sirish L. Shah ." Fault detection and diagnosis in process data using one-class support vector machines" Journal of Process Control 19 (2009) 1627–1639.

[11] M.S. Choudhury, S. Shah, N. Thornhill, D.S. Shook, Automatic detection and quantification of stiction in control valves, Control Engineering Practice 14 (12) (2006) 1395–1412.

[12] N.F. Thornhill, A. Horch, Advances and new directions in plant-wide disturbance detection and diagnosis, Control Engineering Practice 15 (10) (2007) 1196–1206.

[13] Julien Rabatel , Sandra Bringay, Pascal Poncelet." Anomaly detection in monitoring sensor data for preventive maintenance" Expert Systems with Applications 38 (2011) 7003–7015.

[14] X. Huang, H. Qi, X. Liu, Implementation of fault detection and diagnosis system for control systems in thermal power plants, in: Proceedings of the 6th World Congress on Intelligent Control and Automation, Dalian, China, June 21–23, 2006.

Golriz Amooee was born in Tehran, Iran in 1987. She received her B.S. degree with a first class Honors in Information Technology from Islamic Azad University, Parand Branch, Iran, in 2009 and currently she is a M.S. student in the Department of Information Technology at University of Qom, Iran. She specializes in the field of Customer Relationship Management (CRM), Information Security Management and ISO 27001.

Dr Behrouz Minaei-Bidgoli obtained his Ph.D. degree from Michigan State University, East Lansing, Michigan, USA, in the field of Data Mining and Web-Based Educational Systems in Computer Science and Engineering Department. He is working as an assistant professor in Computer Engineering Department of Iran University of Science & Technology, Tehran, Iran. He is also leading at a Data and Text Mining research group in Computer Research Center of Islamic Sciences, NOOR co. Qom, Iran, developing large scale NLP and Text Mining projects for Farsi and Arabic languages.

Malihe Bagheri-Dehnavi was born in Qom, Iran in 1988. She received her B.S. degree in and currently she is a M.S. student in the Department of Information Technology at University of Qom, Iran.

IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011
ISSN (Online): 1694-0814
www.IJCSI.org

A Comprehensive Performance Analysis of Proactive, Reactive and Hybrid MANETs Routing Protocols

Kavita Pandey[1], Abhishek Swaroop[2]
Comp. Sc. Deptt., JIIT, Noida

Abstract

A mobile Ad-hoc network (MANET) is a dynamic multi hop wireless network established by a group of nodes in which there is no central administration. Due to mobility of nodes and dynamic network topology, the routing is one of the most important challenges in ad-hoc networks. Several routing algorithms for MANETs have been proposed by the researchers which have been classified into various categories, however, the most prominent categories are proactive, reactive and hybrid. The performance comparison of routing protocols for MANETs has been presented by other researcher also, however, none of these works considers proactive, reactive and hybrid protocols together. In this paper, the performance of proactive (DSDV), reactive (DSR and AODV) and hybrid (ZRP) routing protocols has been compared. The performance differentials are analyzed on the basis of *throughput, average delay, routing overhead and number of packets dropped* with a variation of number of nodes, pause time and mobility.

Keywords: *MANET, proactive, reactive, hybrid.*

1. Introduction

Mobile Ad-hoc Networks (MANETs) are self configuring networks consisting of mobile nodes that are communicating through wireless links. There is a cooperative engagement of a collection of mobile nodes without the required intervention of any centralized access point or existing infrastructure. The nodes move arbitrarily; therefore, the network may experience unpredictable topology changes. It means that a formed network can be deformed on the fly due to mobility of nodes. Hence, it is said that an ad-hoc wireless network is self organizing and adaptive. Due to infrastructure less and self organizing nature of ad-hoc networks, it has several applications in the area of commercial sector for emergency rescue operations and disaster relief efforts. MANETs also provides a solution in the field of military battlefield to detect movement of enemies as well as for information exchange among military headquarters and so on [1]. Also, MANET provides an enhancement to cellular based mobile network infrastructure. Nowadays, it is an inexpensive alternative for data exchange among cooperative mobile nodes [2].

For communication among two nodes, one node has to check that the receiving node is with in the transmission range of source (Range of a node is defined with the assumption that mobile hosts uses wireless RF transceivers as their network interface), if yes, then they can communicate directly otherwise, with the help of intermediate nodes communication will take place. Each node will act as a host as well as a router. All the nodes should be cooperative so that exchange of information would be successful. This cooperation process is called as routing [3, 4].

Due to the presence of mobility, the routing information will have to be changed to reflect changes in link connectivity. There are several possible paths from source to destination. The routing protocols find a route from source to destination and deliver the packet to correct destination. The performance of MANETs is related to efficiency of the MANETs routing protocols [5] and the efficiency depends on several factors like convergence time after topology changes, bandwidth overhead to enable proper routing, power consumption and capability to handle error rates.

The figure 1 shows the prominent way of classifying MANETs routing protocols. The protocols may be categorized into two types, Proactive and Reactive. Other category of MANET routing protocols which is a combination of both proactive and reactive is referred as Hybrid.

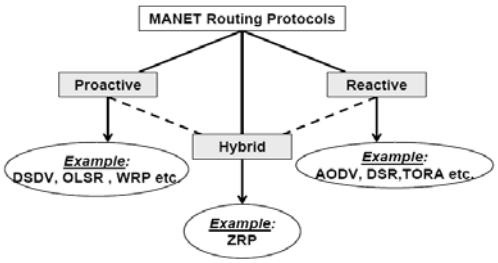


Figure 1 Classification of MANET routing protocols

Proactive routing protocols: In it, all the nodes continuously search for routing information with in a network, so that when a route is needed, the route is already known. If any node wants to send any information to another node, path is known, therefore, latency is low. However, when there is a lot of node movement then the cost of maintaining all topology information is very high [6].

Reactive Routing protocols: Whenever there is a need of a path from any source to destination then a type of query reply dialog does the work [7, 8].Therefore, the latency is high; however, no unnecessary control messages are required.

Hybrid routing protocols: These protocols incorporates the merits of proactive as well as reactive routing protocols. A hybrid routing protocol should use a mixture of both proactive and reactive approaches. Hence, in the recent years, several hybrid routing protocols are proposed like ZRP, ZHLS, SHARP and NAMP etc [7, 9].

In recent years, a variety of routing protocols have been proposed and a comparative analysis of routing protocols has been done either on the basis of simulation results by different simulators like OPNET, NS2, OMNET++ etc. or analytically . In some cases, the comparative analysis is done between reactive routing protocols based on some performance metrics and in other cases between proactive routing protocols. Few researchers have done the simulation based comparison between on demand and table driven routing protocols. The present paper comparatively analyzes all three categories of MANETs routing protocols namely, proactive, reactive and hybrid protocols. In order to compare the protocols, we selected the representative protocols from each category; DSDV from proactive, ZRP from hybrid, and AODV and DSR from the reactive. The performance metrics considered are *throughput*, *average delay*, *routing overhead and number of packets dropped*. The performance differentials are analyzed by varying number of nodes, pause time and mobility using NS2 simulator.

The rest of the paper is organized as follows. The related work has been discussed in section 2. Section 3 provides a brief summary about these protocols. In the section 4, the simulation environment, performance metrics used and results have been discussed. Section 5 concludes the present exposition.

2. Related Work

While most of the work done related to the performance comparison of MANETs routing protocols includes either purely reactive protocols or purely proactive protocols. Some researchers have done a comparative study on reactive and proactive or reactive and hybrid protocols. The table 1 summarizes the work done by various researchers related to performance analysis of MANETs routing protocols.

Table 1: Related work

Author Name Reference	Protocols Used	Simulator	Performance Metrics	Variable Parameters
Guntupalli et al. [5]	DSDV, DSR, AODV	NS2	Average End to End Delay, Normalized Routing Load, Packet Delivery Ratio	Number of Nodes, Speed, Pause time, Transmission Power
Yogesh et al. [10]	AODV, DSR	GLOMOSIM	Packet Delivery Ratio, End to End Delay, Normalized routing overhead	Number of nodes, Speed, Pause time
Chenna et al [11]	DSDV, AODV, DSR, TORA	NS2	Throughput, Routing Overhead, Path Optimality, Packet Loss, Average delay	Traffic Loads, Movement patterns
G. Jayakumar et al. [3]	AODV, DSR	NS2	Packet Delivery Ratio, Normalized Routing Load, MAC load and Average End to End Delay	Number of Sources, Speed, Pause time
Birdar et al.[2]	AODV, DSR	NS2	Packet Delivery Ratio, Routing Overhead, Normalized Routing Overhead and Average End to End Delay	Speed
Kapang et al. [1]	AODV, DSR, DSDV	NS2	Packet Delivery Ratio, Average End to End Delay and Routing Overhead	Pause Time
Vijayalaskhmi et al. [12]	DSDV, AODV	NS2	Packet Delivery Fraction, Average End to End Delay and Throughput	Number of Nodes, Speed, Time
Shaily et al. [13]	AODV, DSR, ZRP	QualNet	TTL based Hop Count, Packet Delivery Ratio and Average End to End Delay	Pause Time
Li Layuan et al. [14]	DSDV, AODV, DSR, TORA	NS2	Average Delay, Jitter, Routing Load, Loss Ratio, Throughput and Connectivity	Network Size

It is evident from table 1 that, no one has presented the comparison of performance differentials among proactive, reactive and hybrid protocols.

3. MANETs Routing Protocols

3.1 DSDV (Destination Sequence Distance Vector)

It is a proactive routing protocol and based on the distributed Bellman-Ford Algorithm. The improvement from distance vector in wired routing protocol is in the terms of avoidance of routing loops. Each node maintains a routing table which has the list of all the possible

destinations and number of routing hops to reach the destination. Whenever some packet comes to node, routing table is to be consulted to find the path. DSDV uses a concept of sequence numbers to distinguish stale routes from new routes and the sequence number is generated by the destination node. To maintain consistency in routing table, DSDV sends routing updates periodically [15]. Therefore, a lot of control message traffic which results in an inefficient utilization of network resources. To overcome this problem, DSDV uses two types of route update packets: *full dump, incremental packets* [16, 17].

3.2 DSR (Dynamic Source Routing)

DSR is a pure reactive routing protocol which is based on the concept of source routing. DSR protocol is composed of two important phases: *route discovery* and *route maintenance.* DSR does not employ any periodic routing advertisement packets, link status sensing or neighbor detection packets [15]. Therefore, the routing packet overhead is less because of its on-demand nature. Every node maintains a route cache to store recently discovered paths. Whenever a route is required for a particular destination then that particular node will consult route cache to determine whether it has already a route to the destination or not. If available route is not expired then that route will be used otherwise a route discovery process is initiated by broadcasting the *route request packet (RREQ)*. When any of the nodes receives RREQ packet, the node will check from their cache or from their neighbors whether it knows a route to the destination. If it does not, the node will add its own address to the route record of the packet and forwards it to their neighbors. Otherwise; a *route reply packet (RREP)* is generated that is unicast back to the original source.

Due to dynamic nature of the environment, any route can fail anytime. Therefore, the route maintenance process will constantly monitors the network and notify the other nodes with the help of route error packets as well as route cache would be updated [16, 18].

3.3 AODV (Ad-hoc On-demand Distance Vector)

AODV algorithm is pure reactive in nature and it contains the properties of both DSR and DSDV protocols. AODV algorithm is an improvement on DSDV in the sense that it minimizes the number of broadcasts. AODV borrows the concept of hop by hop routing, sequence numbers, periodic beacon messages from DSDV protocol [15]. Like DSR, it is on-demand protocol but unlike source routing. When a node wants to send a message to destination node, first it will check whether it has a valid route to the destination or not. If not, then it broadcast a *route request packet* (RREQ) to its neighbors which then forwards the request to their neighbors and so-on, until either it reaches to the intermediate node which has a valid route for the destination or the destination node. AODV uses destination sequence numbers to ensure that it contains most recent information and all routes are loop-free. Once the route request has reached the destination or an intermediate node with a valid route, the destination/intermediate node responds by unicasting a *route reply (RREP) message* back to the neighbor node from which it first received the RREQ [16, 19]. The route maintenance process in AODV is performed with the *route error (RERR) message*. *Hello messages* are used for periodic local broadcast to maintain the local connectivity of the network.

3.4 ZRP (Zone Routing Protocol)

Zone routing protocol is a hybrid protocol. It combines the advantages of both proactive and reactive routing protocols. A routing zone is defined for every node. Each node specifies a zone radius in terms of hops. Zones can be overlapped and size of a zone affects the network performance. The large routing zones are appropriate in situations where route demand is high and / or the network consists of many slowly moving nodes [15]. On the other hand, the smaller routing zones are preferred where demand for routes is less and /or the network consists of a small number of nodes that move fast relative to one another. Proactive routing protocol works with in the zone whereas; reactive routing protocol works between the zones.

ZRP consists of three components: **1) the proactive Intra zone routing protocol (IARP), 2) the reactive Inter zone routing protoc ol (IERP) and 3) Bordercast resolution protocol (BRP).** Each component works independently of the other and they may use different technologies in order to maximize efficiency in their particular area. The main role of IARP is to ensure that every node with in the zone has a consistent updated routing table that has the information of route to all the destination nodes with in the network. The work of IERP gets started when destination is not available with in the zone. It relies on bordercast resolution protocol in the sense that border nodes will perform on-demand routing to search for routing information to nodes residing outside the source node zone [20].

4. Simulation

There are several simulators available like OMNET++, QualNet, OPNET and NS2. Here, NS2 is used for simulation experiments since it is preferred by the networking research community. NS2 is an object oriented simulator, written in C++ and OTcl (Object oriented Tool command language) as the frontend. If the components have to be developed then both Tcl (Tool command language – scripting language) and C++ have to be used. In this section, we have described about the performance metrics and implementation details of all four

MANETs routing protocols namely, DSDV, DSR, AODV and ZRP.

4.1 Performance Metrics

The following performance metrics are considered for evaluation of MANETs routing protocols:

Throughput: the ratio of data packets received to the destination to those generated by source.

Average Delay: it includes all possible delays caused by buffering during route discovery latency, queuing at the interface queue, retransmission delays at the MAC, and propagation and transfer times. It is the average amount of time taken by a packet to go from source to destination. [19]

No. of packets dropped: it is the number of packets lost by routers at the network layer due to the capacity of buffer or the packet buffering time exceeds the time limit.

Routing Overhead: it is the number of routing packets which would be sent for route discovery and maintenance.

All the above mentioned performance metrics are quantitatively measured. For a good routing protocol, throughput should be high where as other three parameters value should be less.

4.2 Implementation

The simulation parameters considered for the performance comparison of MANETs routing protocols are given below:

Table 2: Simulation Setup parameters

Platform	Linux, Fedora core 9
NS Version	ns-allinone-2.34
Protocol	DSDV, AODV, DSR , ZRP
Mobility Model	Random way Point
Area	500 * 500 m
Experiment Duration	150 sec
Traffic Type	CBR
Radio Propagation	TwoRayGround
MAC layer Protocol	Mac/802_11
Packet size	512 bytes
Antenna type	Antenna/OmniAntenna
Number of nodes	10, 20, 30, 40, 50
Maximum Speed	5, 10, 15, 20, 25 m/s
Pause time	10, 50, 100, 150, 200

NS2 provides the implementation of DSDV, AODV and DSR protocols. However, for ZRP, a patch has been integrated into NS2 package [21, 22]. The Tcl code has been written to set up the network components which includes the parameters defined in Table 2. For **traffic model**, cbrgen utility has been used which creates CBR and TCP traffic connections between nodes [23]. The different traffic files have been generated by varying the number of nodes with CBR traffic source at a rate of 4 packets / sec and keeping maximum number of connections as 20 to 40.

For **mobility model**, setdest utility has been used to create node positions and their movements [23]. In order to perform simulation experiments, twenty five different scenario files have been generated by varying the number of nodes and pause time and keeping other values constant. Other twenty five scenario files have been generated by varying the number of nodes and maximum speed by keeping the pause value as 2 seconds. Pause time, Max speed and number of nodes are varied according to the table 2.

The **Tcl** file generates different trace files according to different MANETs routing protocols. In order to test the behavior of different protocols, the trace files have been parsed with the help of programs written in **Python** language to extract the information needed to measure the performance metrics. After getting the values of different performance metrics according to different routing protocols, **XGraph** utility is used to plot the graphs. **Network Animator (NAM)** is used to graphically visualize the simulation [24, 25,26].

4.3 Results

Simulation Results have been presented in the group of four figures where each figure is corresponding to one performance metric. The performance metrics considered are *throughput, average delay*, *number of packets dropped* and *routing overhead*. In all graphs, x-axis specifies the number of nodes and y-axis indicates the value of performance parameter.

We have presented the analysis of results according to different performance metrics. With each performance metric, results are analyzed by changing the number of nodes, speed and pause time. The group of graphs numbered from 1 to 5 shows the simulation results by varying the pause values and number of nodes, however, the maximum speed is kept as constant that is 2 m/s. Whereas, the next group of graphs numbered from 6 to 10 shows the simulation results by keeping the pause value constant i.e., 2 seconds and varying the maximum speed and number of nodes. The minimum speed is taken as 1 m/s, so, that nodes will move with an average speed.

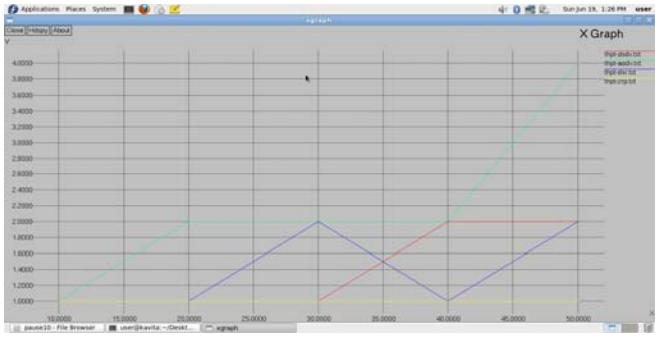
Throughput: It is evident from the results that throughput of AODV is better as compared to other protocols. Moreover, the change in the pause value does not have any effect on AODV performance. Generally, for all the protocols, by increasing the number of nodes, throughput also increases. In DSDV, initially (before the convergence of roots), some packets are sent and get dropped, therefore, it has low throughput as compared to AODV and DSR. With pause value 200 and number of nodes 10, the throughput of DSDV is zero. The throughput of ZRP does not change even on changing the pause value or speed or the number of nodes. The reason behind this

phenomenon may be the fixed zone radius. On changing the pause value, the throughput of DSR has an oscillating behavior. One possible reason is that DSR uses route cache and the routes stored in the cache might become stale after sometime. However, by increasing the speed, the throughput of DSR decreases. This can be due to the mobility of nodes which may increase the chances of path failure.

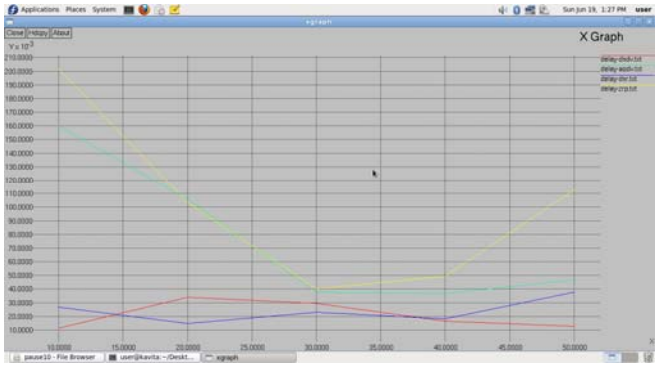
Average Delay: From the graphs related to average delay, it can be seen that AODV and ZRP has higher average delay where as other two protocols experiences less delay. In DSR, due to the caching of routes, the average delay is reduced. However, as the number of nodes increases, DSR exhibits significantly higher delay then other three protocols. This may be due to the increasing node density because of which the number of data sessions increases which leads to increased end to end delay. Average delay of DSDV is less in comparison to other three protocols since it is a proactive protocol. The routes for all the destinations are maintained in routing tables. When speed increases, there is no effect on average delay. However, as the number of nodes increases, the delay increases due to time consumed in computation of routes, however, once routes have been created; the delay becomes less as evident from the graphs. Since, AODV is a reactive protocol, the routes are created on demand; therefore, it experiences a higher delay. As speed increases, spikes in AODV are higher; however, for higher mobility, AODV has less delay as compared to previous values. In case of ZRP, initially, when number of nodes is less, it has a higher delay because of the route creation and table maintenance, then the delay decreases and after that it gives a mediocre performance which is expected because of its hybrid nature. In ZRP, on increasing the speed and the number of nodes, the delay increases because of difficulty in setting routes due to contention and high mobility.

Number of Dropped Packets: In DSDV protocol, more number of packets gets dropped as compared to other protocols. Generally, the value is higher when the number of nodes is less. The reason may be sending the data packets before convergence of routes. DSR and AODV experience a similar behavior that dropped packets are less, which specifies their high reliability. However, in DSR the number of dropped packets is marginally less in comparison to AODV. This reduction may be due to the fact that DSR maintains route cache. Generally, by increasing the pause value, number of dropped packets also increases. Initially, ZRP has less number of dropped packets; however, as number of nodes increases, there is a sharp increase in the value. In every protocol, the number of dropped packets increases on increasing the speed due to difficulty in path creation. With an increase in speed, descending order of performance corresponding to number of dropped packets is ZRP, DSR, AODV, and DSDV. On increasing the speed, the number of dropped packets in DSDV protocol is high since it is a table-driven routing protocol.

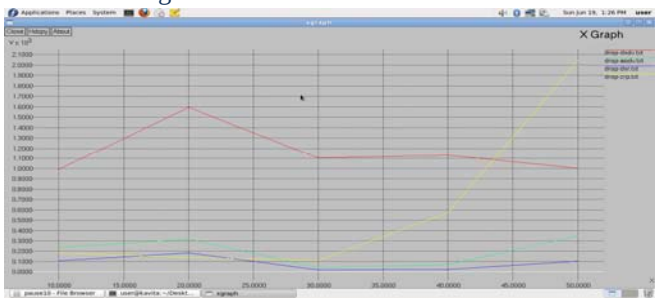
Routing Overhead: ZRP and AODV have more routing overhead in comparison to DSR and DSDV routing protocols. In DSR, the routes are maintained only between the nodes those want to communicate as well as a single route discovery may yield many routes to the destination, therefore, the routing overhead is less. Where as, in DSDV, the concept of table maintenance reduces the routing overhead. In ZRP, maintaining the zone radius as well as electing the border nodes and switching from proactive to reactive or vice-versa, more number of control packets are needed. As number of nodes increases, the routing overhead increases because of increasing node density. In ZRP and AODV, routing overhead increases by a large amount where as, in DSDV and DSR, it increases marginally. In DSDV and DSR, there is not a significant effect of pause value or speed value. Where as, in ZRP overhead is reduced by a small amount with respect to increase in pause value and speed. This scenario is reversed in AODV protocol, that is by increasing the speed and pause value, overhead also increases. At the last, it can be concluded that the routing overhead increases with increasing number of nodes; however, a change in pause value or speed does not adversely affect the performance.



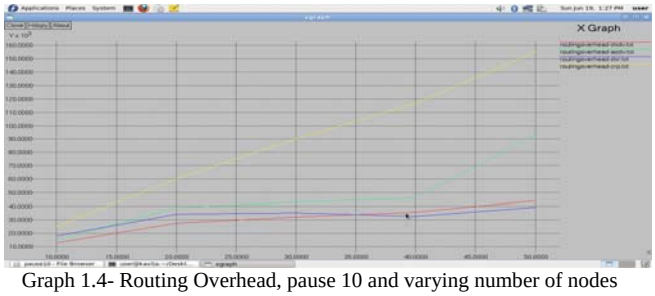
Graph 1.1- Throughput, pause 10 and varying number of nodes



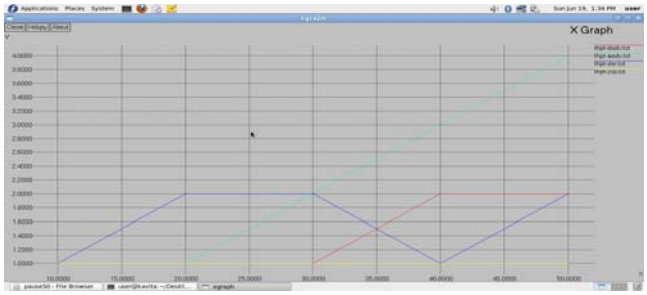
Graph 1.2- Average Delay, pause 10 and varying number of nodes



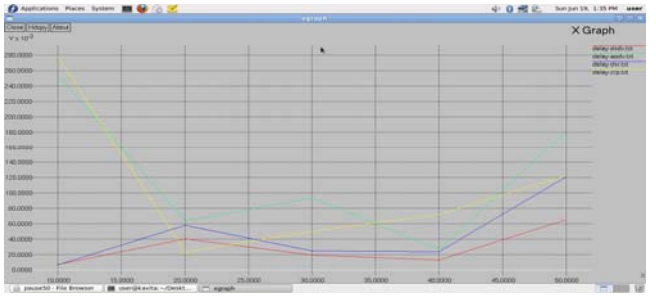
Graph 1.3- Dropped Packets, pause 10 and varying number of nodes



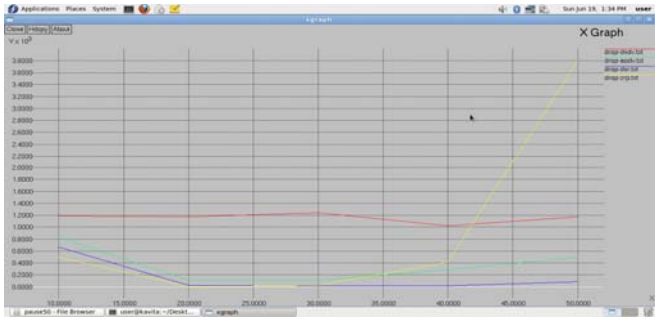
Graph 1.4- Routing Overhead, pause 10 and varying number of nodes



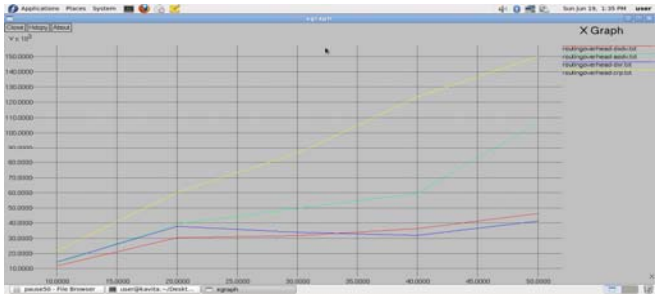
Graph 2.1- Throughput, pause 50 and varying number of nodes



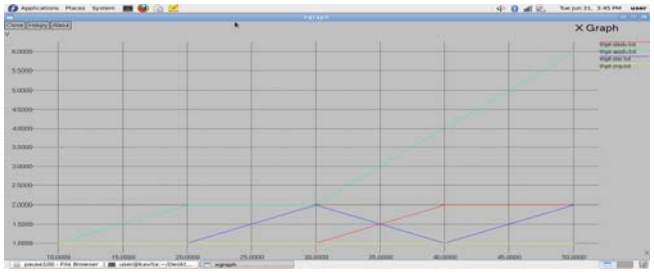
Graph 2.2- Average Delay, pause 50 and varying number of nodes



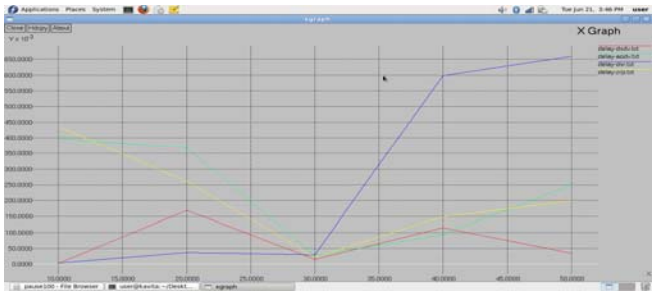
Graph 2.3- Dropped Packets, pause 50 and varying number of nodes



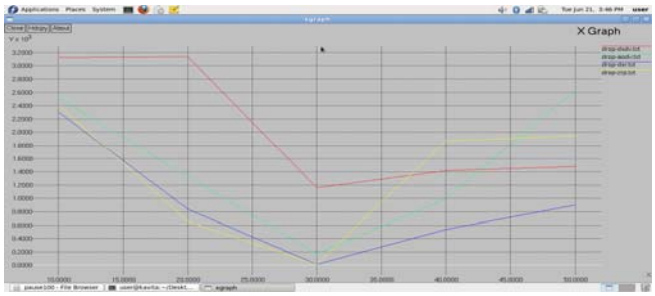
Graph 2.4- routing Overhead, pause 50 and varying number of nodes



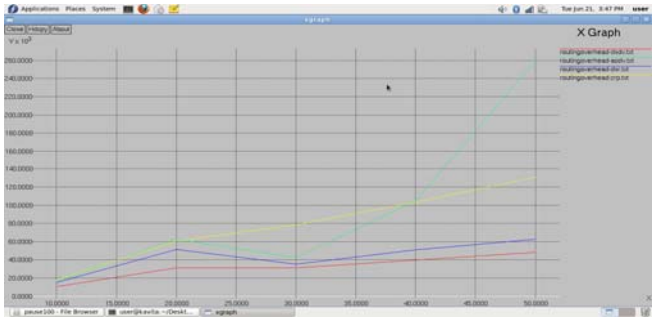
Graph 3.1- Throughput, pause 100 and varying number of nodes



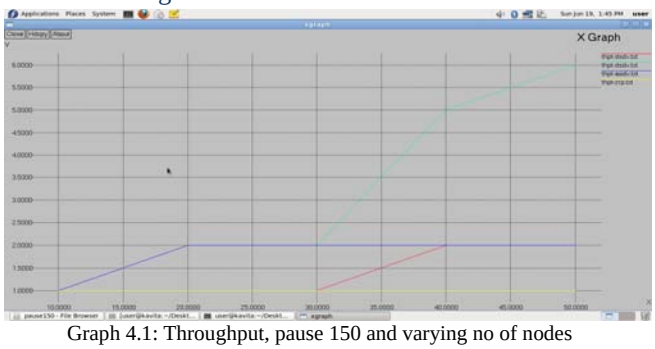
Graph 3.2- Average Delay, pause 100 and varying number of nodes



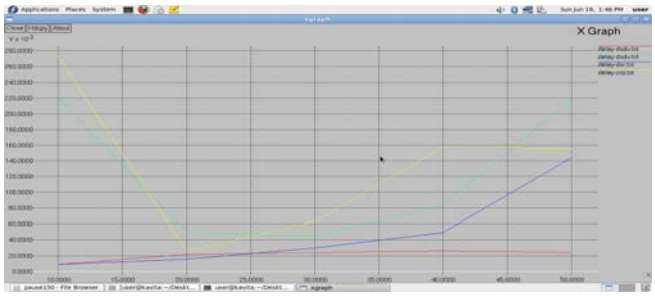
Graph 3.3- Dropped packets, pause 100 and varying number of nodes



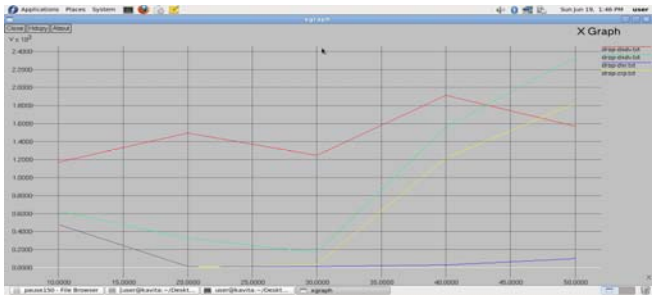
Graph 3.4- Routing Overhead, pause 100 and varying number of nodes



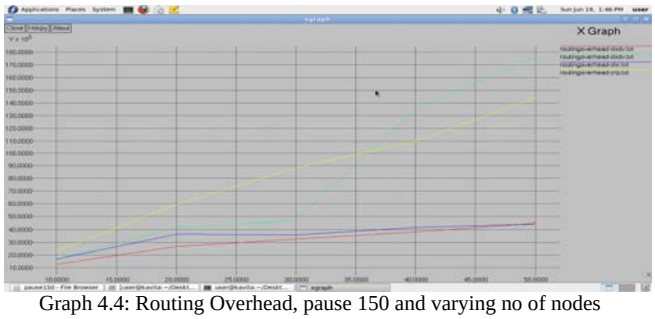
Graph 4.1: Throughput, pause 150 and varying no of nodes



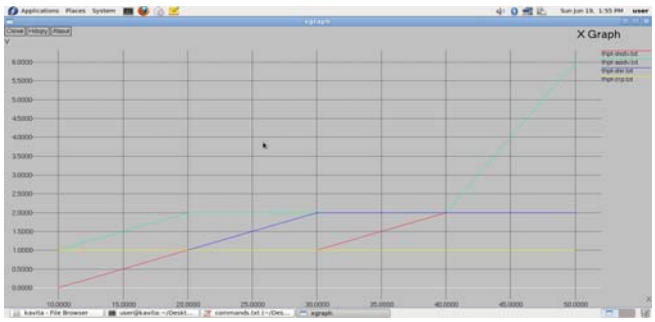
Graph 4.2: Average Delay, pause 150 and varying no of nodes



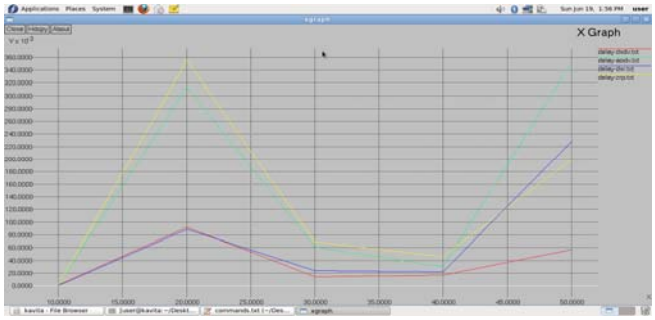
Graph 4.3: Dropped Packets, pause 150 and varying no of nodes



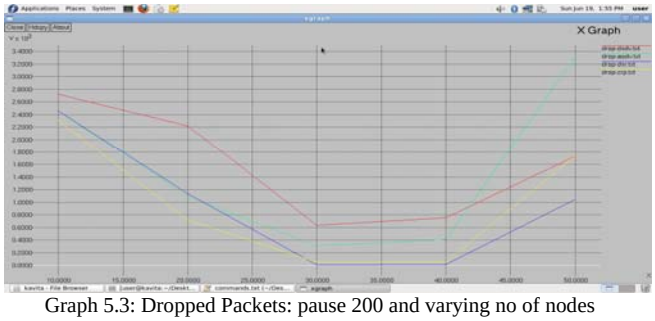
Graph 4.4: Routing Overhead, pause 150 and varying no of nodes



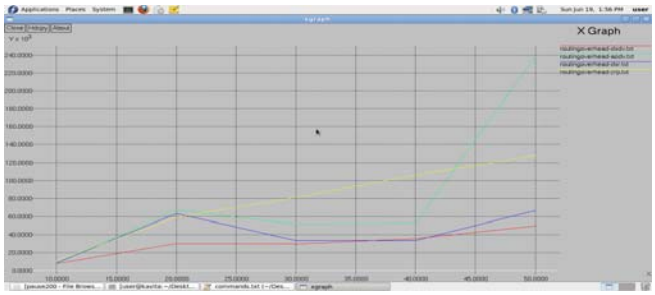
Graph 5.1: Throughput: pause 200 and varying no of nodes



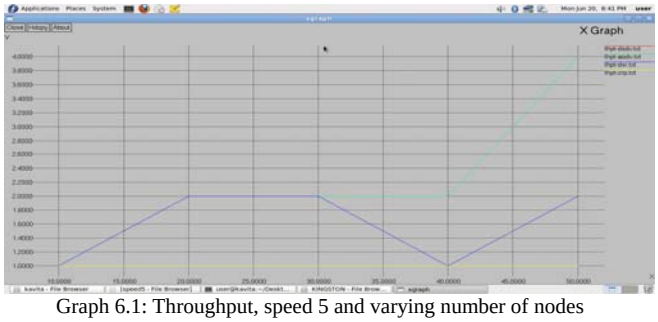
Graph 5.2: Average Delay: pause 200 and varying no of nodes



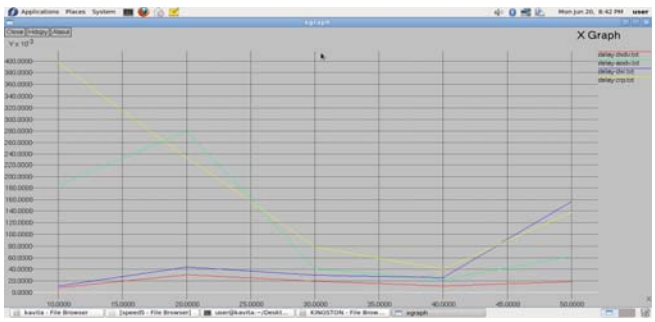
Graph 5.3: Dropped Packets: pause 200 and varying no of nodes



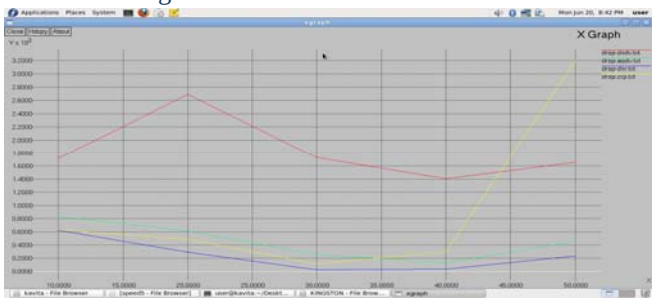
Graph 5.4: Routing Overhead: pause 200 and varying no of nodes



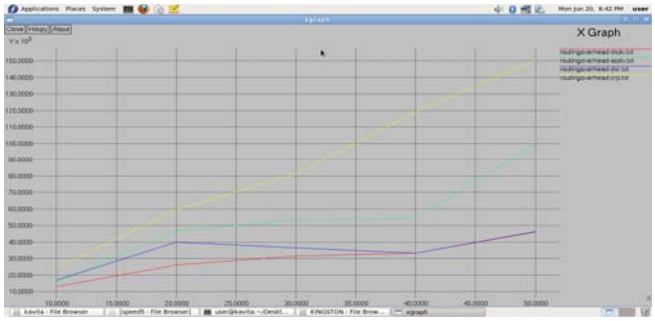
Graph 6.1: Throughput, speed 5 and varying number of nodes



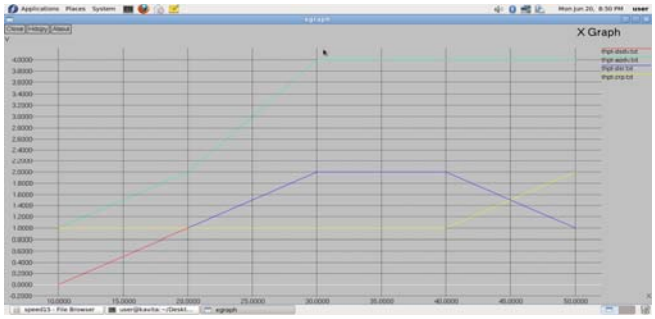
Graph 6.2: Average Delay, speed 5 and varying number of nodes



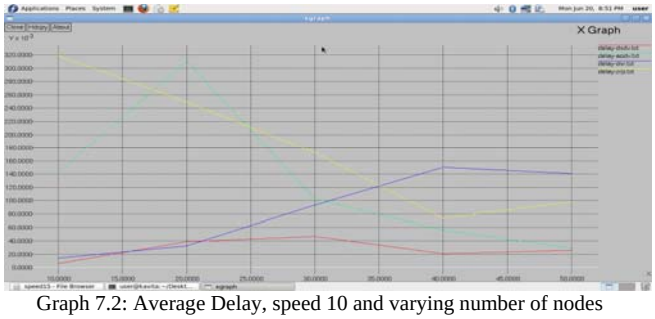
Graph 6.3: Dropped Packets, speed 5 and varying number of nodes



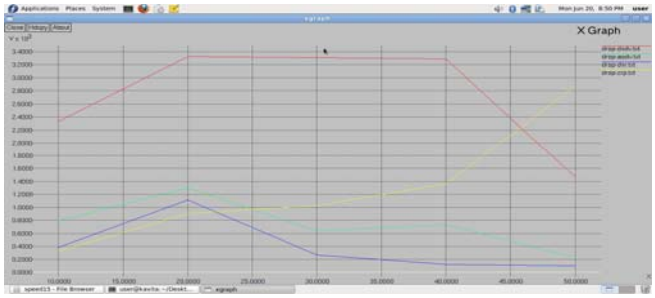
Graph 6.4: Routing Overhead, speed 5 and varying number of nodes



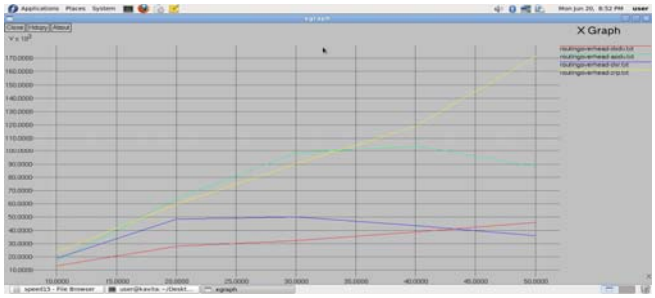
Graph 7.1: Throughput, speed 10 and varying number of nodes



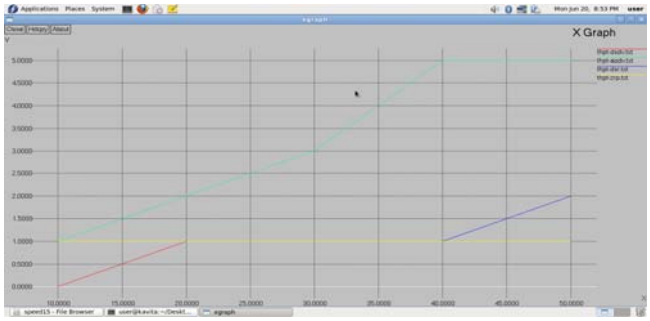
Graph 7.2: Average Delay, speed 10 and varying number of nodes



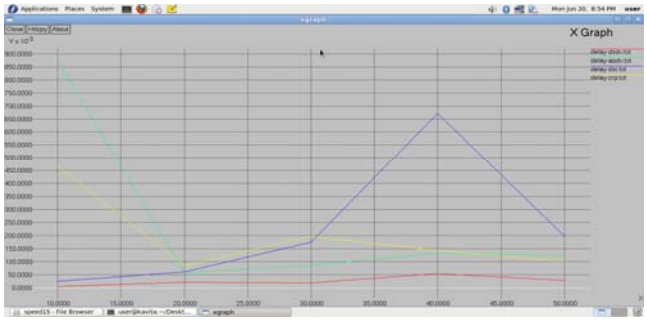
Graph 7.3: Dropped Packets, speed 10 and varying number of nodes



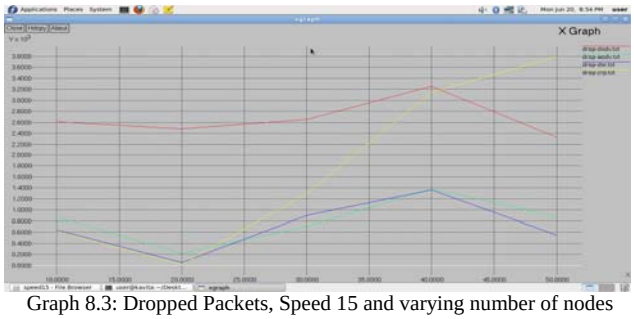
Graph 7.4: Routing Overhead, speed 10 and varying number of nodes



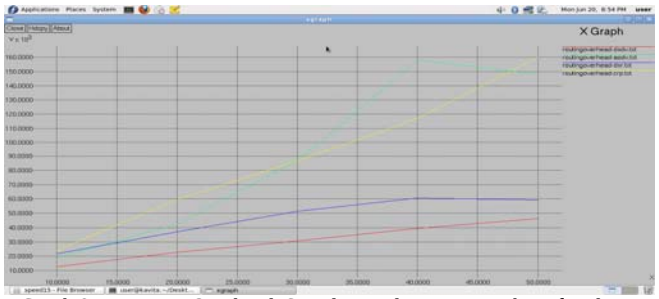
Graph 8.1: Throughput, Speed 15 and varying number of nodes



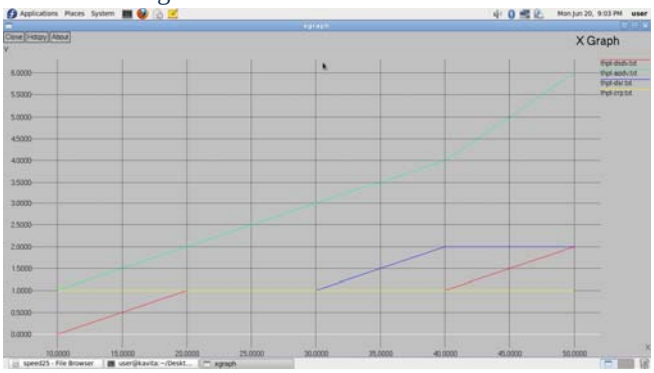
Graph 8.2: Average Delay, Speed 15 and varying number of nodes



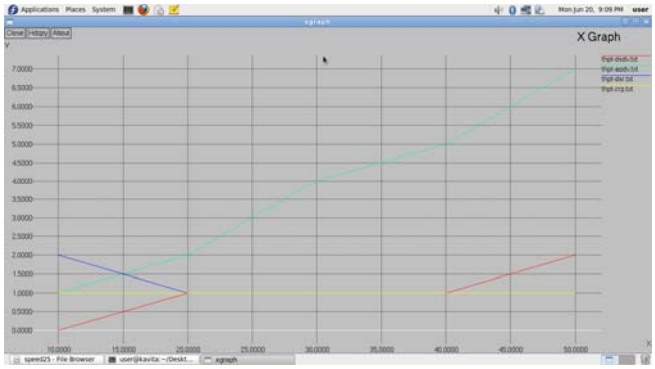
Graph 8.3: Dropped Packets, Speed 15 and varying number of nodes



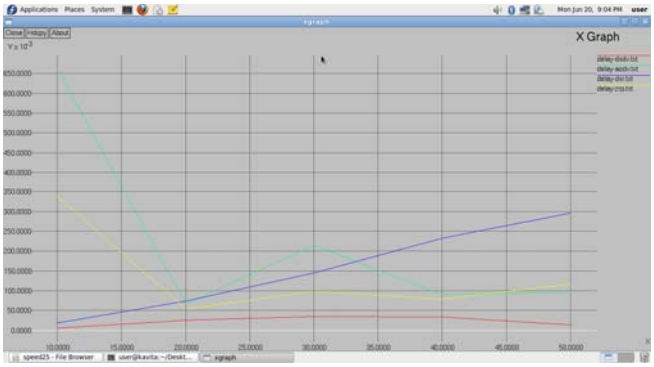
Graph 8.4: Routing Overhead, Speed 15 and varying number of nodes



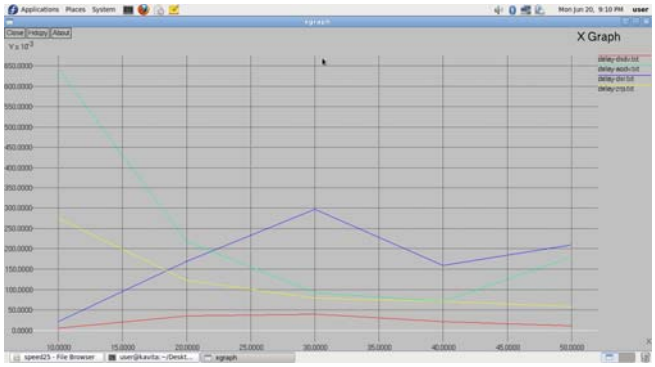
Graph 9.1: Throughput, Speed 20 and varying number of nodes



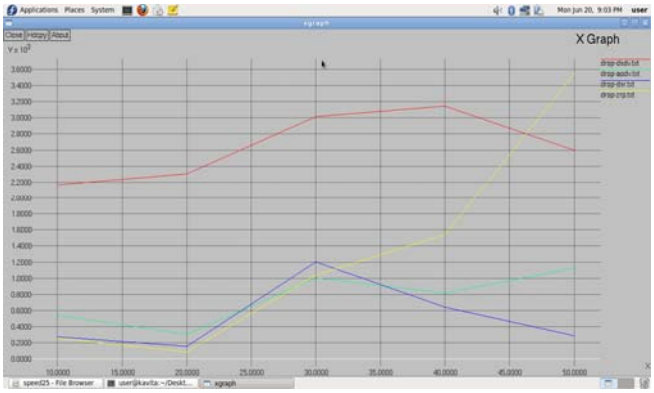
Graph 10.1: Throughput, Speed 25 and varying number of nodes



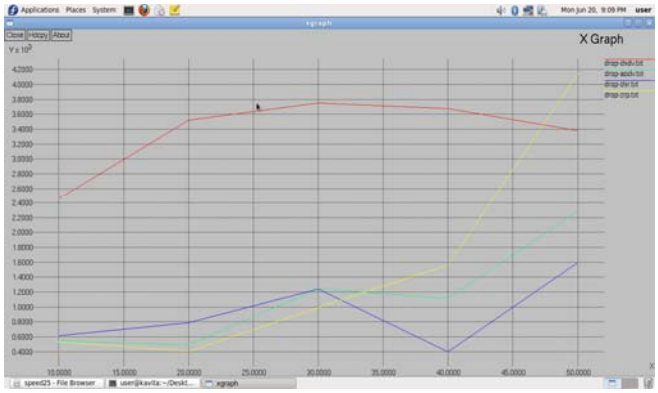
Graph 9.2: Average Delay, Speed 20 and varying number of nodes



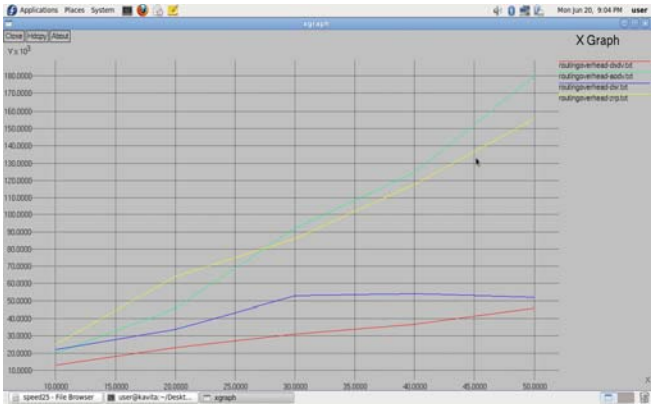
Graph 10.2: Average Delay, Speed 25 and varying number of nodes



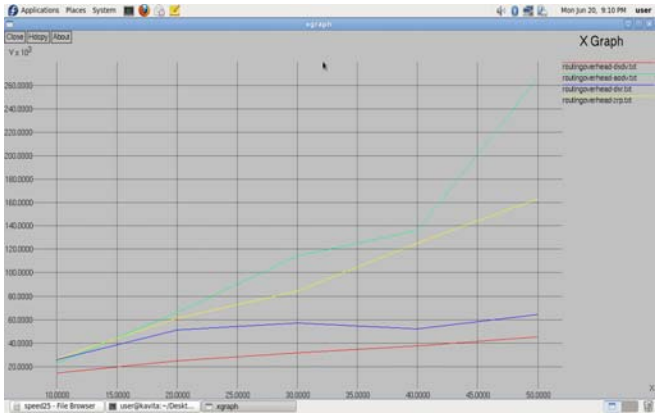
Graph 9.3: Dropped Packets, Speed 20 and varying number of nodes



Graph 10.3: Dropped Packets, Speed 25 and varying number of nodes



Graph 9.4: Routing Overhead, Speed 20 and varying number of nodes



Graph 10.4: Routing Overhead, Speed 25 and varying number of nodes

5. Conclusion

In the present exposition, the performance of MANET routing protocols is examined with respect to following four performance metrics namely, throughput, average delay, number of packets dropped and routing overhead. DSDV is a proactive protocol, where as, AODV and DSR falls under the category of reactive protocol and ZRP is a hybrid protocol. The simulation results suggest that each protocol performs well in some scenario yet has some drawbacks in other cases. In terms of throughput, AODV performance is better than others whereas, DSDV performance poorly sometimes. Another disadvantage of DSDV is that the number of dropped packets is also significantly higher. ZRP throughput does not change even with a change in mobility or pause time because of its hybrid nature. The performance of DSR is good in terms of routing overhead and number of packets dropped due to route cache. It can also be concluded from the simulation results that the reliability of AODV and DSR protocols is better than other two protocols.

6. References

[1]. Kapang Lego, Pranav Kumar Singh, Dipankar Sutradhar, *"Comparative Study of Adhoc Routing Protocol AODV, DSR and DSDV in Mobile Adhoc NETwork"*, Indian Journal of Computer Science and Engineering Vol. 1 No. 4 364-371, 2011.

[2]. S.R. Birdar, Hiren H D Sarma, Kalpana Sharma, Subir Kumar Sarkar , Puttamadappa C, *Performance Comparison of Reactive Routing Protocols of MANETs using Group Mobility Model"*, International Conference on Signal Processing Systems, 2009.

[3]. G. Jayakumar and G. Gopinath, *"Performance comparison of two on-demand routing protocols for ad-hoc networks based on random way point mobility model,"* American Journal of Applied Sciences, vol. 5, no. 6, pp. 649-664, June 2008.

[4]. S. Ahmed and M. S. Alam, *"Performance Evaluation of important ad hoc networks protocols"*, EURASIP Journal on wireless Communications and networking, Vol: 2006, Article ID 78645, PP 1-11, 2006.

[5].Guntupalli Lakshmikant, A Gaiwak, P.D. Vyavahare, *"Simulation Based Comparative Performance Analysis of Adhoc Routing Protocols"*, in proceedings of TENCON 2008.

[6]. OLSR, internet draft, http://tools.ietf.org/html/draft-ietf-manet-olsr-00

[7]. G. Vijaya Kumar, Y. Vasudeva Reddyr, M. Nagendra, "*Current Research Work on Routing Protocols for MANET: A Literature Survey*", International Journal on Computer Science and Engineering, Vol. 02, No. 03, pp. 706-713, 2010.

[8]. Vincent D. Park, M. Scott Corson, Temporally-Ordered Routing Algorithm (TORA) version 1: functional specification, Internet-Draft, draft-ietf-manet-tora-spec- 00.txt, November 1997.

[9]. V. Ramasubramanian, Z. J. Haas, and E. G. Sirer, "SHARP: A Hybrid Adaptive Routing Protocol for Mobile Ad Hoc Networks," The Fourth ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc), pp. 303-314, 2003.

[10]. Yogesh Chaba, Yudhvir Singh, Manish Joon, *"Simulation Based Performance Analysis of On-Demand Routing Protocols in MANETs,"* Second International Conference on Computer Modeling and Simulation, 2010.

[11]. Chenna Reddy, P.; ChandraSekhar Reddy, P., *"Performance Analysis of Adhoc Network Routing Protocols,"* ISAUHC '06, International Symposium on Ad Hoc and Ubiquitous Computing, vol., no., pp.186-187, 20-23 Dec. 2006.

[12]. Vijayalaskhmi M. Avinash Patel, Linganagouda Kulkarni, *"QoS Parameter Analysis on AODV and DSDV Protocols in a Wireless Network"*, International Journal of Communication Network and Security, Volume-1, Issue-1, 2011.

[13]. Shaily Mittal, Prabhjot Kaur, "*Performance Comparison of AODV, DSR and ZRP Routing Protocols in MANET's"*, International Conference on Advances in Computing, Control, and Telecommunication Technologies, 2009.

[14]. Li Layuan, Li Chunlin, Yaun Peiyan, "Performance evaluation and simulation of routing protocols in ad hoc networks", Computer Communications 30 (2007) 1890-1898.

[15]. C.K. Toh, Ad Hoc Mobile Wireless Networks Protocols and Systems, Pearson Education, 2007.

[16]. Sunil Taneja, Ashwani Kush, *"A Survey of Routing Protocols in Mobile Adhoc Networks"*, International Journal of Innovation, Management and Technology, Vol. 1, No. 3, August 2010.

[17]. Charles E. Perkins and Pravin Bhagwat, "Highly Dynamic Destination-Sequenced Distance-Vector routing (dsdv) for Mobile Computers", 1994.

[18]. DSR, internet draft, http://tools.ietf.org/html/draft-ietf-manet-dsr-10 .

[19]. AODV, internet draft, http://tools.ietf.org/html/draft-ietf-manet-aodv-09 .

[20]. ZRP, internet draft, http://tools.ietf.org/id/draft-ietf-manet-zone-zrp-04.txt .

[21]. ZRP patch, http://magnet.daiict.ac.in/magnet_members/MTech/2007/PatelBrijesh/Simulation.html#Sec_2.

[22]. ZRP Agent Implementation documentation, http://magnet.daiict.ac.in/magnet_members/MTech/2007/PatelBrijesh/Thesis_files/MyZRP/ZRPManual.pdf .

[23]. Yinfei Pan, "*Design Routing Protocol Performance Comparison in NS2: AODV Comparing to DSR as Example"*, Deptt of CS, SUNY Binghamton, Vestal NY 13850.

[24].NS2 Trace format - http://nsnam.isi.edu/nsnam/index.php/NS-2_Trace_Formats .

[25].The ns Manual (formerly ns Notes and Documentation) by Kevin Fall, Kannan Varadhan. http://www.isi.edu/nsnam/ns/doc/ns_doc.pdf

[26]. NS Simulator for beginners, http://www-sop.inria.fr/members/Eitan.Altman/COURS-NS/n3.pdf.

Kavita Pandey She did B.Tech in Computer Science from M.D. University in year 2002 and M.Tech. in Computer Science from Banasthali Vidyapith University in year 2003. She is pursuing PhD from JIIT,Noida. She is working as a Senior Lecturer in JIIT,Noida. Her current research interests include Adhoc Networks, Optimization Techniques and Network Security.

Abhishek Swaroop He received his B.Tech in computer science and engineering from G.B.Pant Univ. Pantnagar in 1992, M.Tech from Punajbi Univ. Patiala in 2004 and Ph.D. in computer engineering from N.I.T. Kurukshetra in 2011. He is currently working as an Assitant professor in JIIT, Noida. His research interests include group mutual exclusion, fault tolerance, MANETs, Sensor networks, Multi core architecture.

IJCSI

ISSN (online): 1694-0814

Please consider to contribute to and/or forward to the appropriate groups the following opportunity to submit and publish original scientific results.

CALL FOR PAPERS International Journal of Computer Science Issues (IJCSI)
Volume 9, Issue 2 – March 2012 Issue

The topics suggested by this issue can be discussed in term of concepts, surveys, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.
See authors guide for manuscript preparation and submission guidelines.

Indexed by Google Scholar, DBLP, CiteSeerX, Directory for Open Access Journal (DOAJ), Bielefeld Academic Search Engine (BASE), SCIRUS, Cornell University Library, ScientificCommons, EBSCO, ProQuest and more.

Deadline: 31st January 2012
Notification: 29th February 2012
Revision: 10th March 2012
Publication: 31st March 2012

Context-aware systems
Networking technologies
Security in network, systems, and applications
Evolutionary computation
Industrial systems
Evolutionary computation
Autonomic and autonomous systems
Bio-technologies
Knowledge data systems
Mobile and distance education
Intelligent techniques, logics and systems
Knowledge processing
Information technologies
Internet and web technologies
Digital information processing
Cognitive science and knowledge
Agent-based systems
Mobility and multimedia systems
Systems performance
Networking and telecommunications
Software development and deployment
Knowledge virtualization
Systems and networks on the chip
Knowledge for global defense
Information Systems [IS]
IPv6 Today - Technology and deployment
Modeling
Software Engineering
Optimization
Complexity
Natural Language Processing
Speech Synthesis
Data Mining

For more topics, please see http://www.ijcsi.org/call-for-papers.php

For more information, please visit the journal website (www.IJCSI.org)

www.IJCSI.org

IJCSI

The International Journal of Computer Science Issues (IJCSI) is a well-established and notable venue for publishing high quality research papers as recognized by various universities and international professional bodies. IJCSI is a refereed open access international journal for publishing scientific papers in all areas of computer science research. The purpose of establishing IJCSI is to provide assistance in the development of science, fast operative publication and storage of materials and results of scientific researches and representation of the scientific conception of the society.

It also provides a venue for researchers, students and professionals to submit ongoing research and developments in these areas. Authors are encouraged to contribute to the journal by submitting articles that illustrate new research results, projects, surveying works and industrial experiences that describe significant advances in field of computer science.

Indexing of IJCSI

1. Google Scholar
2. Bielefeld Academic Search Engine (BASE)
3. CiteSeerX
4. SCIRUS
5. Docstoc
6. Scribd
7. Cornell's University Library
8. SciRate
9. ScientificCommons
10. DBLP
11. EBSCO
12. ProQuest