Journal of Computers

ISSN 1796-203X

Volume 5, Number 5, May 2010

Contents

REGULAR PAPERS

Handwritten Nushu Character Recognition Based on Hidden Markov Model

Jiangqing Wang, Rongbo Zhu
College of Computer Science, South-Central University for Nationalities, Wuhan 430074, China
Email: wjqing2000@yahoo.com.cn

***Abstract*—This paper proposes a statistical-structural character learning algorithm based on hidden Markov model for handwritten Nushu character recognition. The stroke relationships of a Nushu character reflect its structure, which can be statistically represented by the hidden markov model. Based on the prior knowledge of character structures, we design an adaptive statistical-structural character learning algorithm that accounts for the most important stroke relationships, which aims to improve the recognition rate by adapting selecting correct character to the current handwritten character condition. We penalize the structurally mismatched stroke relationships using the prior clique potentials and derive the likelihood clique potentials from Gaussian mixture models. Theoretic analysis proves the convergence of the proposed algorithm. The experimental results show that the proposed method successfully detected and reflected the stroke relationships that seemed intuitively important. And the overall recognition rate is 93.7 percent, which confirms the effectiveness of the proposed methods.**

***Index Terms*—character recognition, statistical-structural learning algorithm, Nushu character, hidden Markov models**

I. Introduction

Nushu (female scripts) was derived from square Chinese characters, and were variations of the later [1]. Nushu was popular in the valley of the Xiaoshui River in Jiangyong County of Hunan Province, and is still used by some senile women nowadays. Researches show that Nushu had more than 1,000 characters, among which 80% were created based on Chinese characters, and only 20% were coinages with unknown origin. Its characters took the shape of rhombus, and were higher on the right part and lower on the left part. They are slender and beautiful, look like Jiaguwen (scripts on tortoise shells and animal bones) at first glance, and retain much familiar trace of Chinese characters [2]. Nushu is composed of very strange characters, which feature strange shapes, a strange way of marking, strange social functions and history. Different from Chinese ideographic characters, Nushu characters are ideographic characters that have a single syllable and indicate their sound. Nushu was the tool of cultural communication for local countryside women, especially middle- and old-aged women. It played its unique social function, and was basically used to create women's works and record women's songs. Nushu works normally were written on delicately made manuscripts, fans, handkerchiefs and pieces of paper. Nushu has academic value from the perspectives of philology, linguistics, sociology, ethnography, and history, etc. Therefore, it is reputed as a wonderful discovery and a wonder in the history of Chinese characters by scholars home and abroad.

A. Related Work

Since more than 80% Nushu characters were created based on Chinese characters, the original Nushu material was handwritten. Therefore, the research scheme of handwritten Nushu character recognition can adopt the scheme of Chinese characters recognition, which is the most common way to. The Chinese character structure is hierarchical: many straight-line strokes constitute independent radicals, which in turn constitute characters [3, 4]. The statistical recognizer extracts the information of the character image into a feature vector by a feature extraction process. The feature vector does not represent the pen-trajectories directly. Generally, the recognizer represents and analyzes the character image by one of the two kinds of method, the statistical method and the structural method [5, 6]. However, it represents the property of the character image, reflecting the character structure indirectly. With such a representation, the statistical recognizer models and analyzes the character using various kinds of statistical methodologies. Wu et al. [7] projected a two–dimensional (2-D) character image along x and y directions. The key features for coarse classification are the Fourier coefficients of the projected profiles. Tseng et al. [8] selected contour direction and crossing count to be the features of their coarse classification method. Chang and Wang [9], inspired by Dr. W. Yun-Wu, used the peripheral shape coding technique to preclassify handwritten Chinese characters. They used ten categories of stroke patterns to code the four corners of a character. However, such systems focused only on the relationship between near or connected stroke pairs. As the result, they were difficult to represent the relationship between the strokes far from each other. Moreover, they were not effective to represent the relationship between more than two strokes because the interstroke feature is difficult to define for more than

Manuscript received January 1, 2009; revised Feberary 15, 2009; accepted Feberary 20, 2009.
Corresponding author, J. Wang, E-mail: wjqing2000@yahoo.com.cn

doi:10.4304/jcp.5.5.663-670

two strokes. Also, they had a problem in combining the stroke matching scores with the matching scores of the stroke relationships. They computed the overall matching score by simply accumulating or multiplying all stroke matching scores and matching scores of the stroke relationships. As the result, the information about the stroke relationship was duplicated because the matching scores of individual strokes also reflect some information about the stroke relationships.

Structure approach has high tolerance to non-structure distortions, such as noise and writing style variations. Since Chinese characters are composed of strokes formed by line segments, most approaches use the geometrical and topological features of strokes as the recognition basis. In order to represent finer information, Liu et al. and Zhang and Xia categorized the strokes into several types, such as horizontal, vertical, slash, back-slash, dot, tick, and hook [10], [11]. The character model was composed of a set of model strokes, each of which belongs to one of the predefined types. By assigning different attributes to each of the stroke types, they represented the properties of the stroke more accurately. However, the performance of such systems heavily depends on the developer’s knowledge because the character models were not systematically trained but manually designed. Kim and Kim represented the stroke by the distributions of its position, slope, and length [12]. Such a statistical modeling is more systematic than the previous methods. Although the character structure was manually specified as was in the previous methods, the position and the shape of each stroke were statistically modeled. Their distributions were estimated from training samples. The statistical stroke modeling is desirable to tolerate writing variation and more robust than the heuristic-based method. Character recognition proceeds by finding the best structural match between the input strokes and the stroke models. Compared with the statistical method, the structural method extracts feature points and line segments from character images and represents their spatial relationships by a relational graph, in which the node denotes the feature point or line segment, and the edge between two nodes denotes their relationships (for example, constraint graph model [10], attributed relational graph [11], and hierarchical random graph [12]). Despite the excellent descriptive ability for fine details of character structures, there are two major problems yet to be solved. The first is the stroke extraction problem—because the strokes are often ambiguous and degraded how to extract the stable ones for modeling their spatial relationships. This problem becomes much more difficult if the thinning preprocessing techniques cause junction-distortions in character skeletons [13]. The second problem lies in that the structural method usually depends heavily on developer’s heuristic knowledge [14, 15], leading to neither the rigorous matching algorithm nor the automatic leaning scheme from training samples.

B. Motivation

Chinese character, as well as Nushu character recognition is admitted as a very difficult problem in character recognition due to (1) very large character set, (2) high complexity of Chinese characters and (3) many similar character patterns. Since both statistical scheme and structural scheme have their merit and demerit. Therefore, a hybrid statistical-structural method is necessary for modeling character structures and recognizes characters. Our approach can be considered to be a convergence between these two threads of research. However, it improves the performance on both sides in term of overall recognition rate.

In this paper, we concentrate on the handwritten Nushu character recognition problem where few research works have done. A statistical-structural character learning algorithm based on hidden Markov model is proposed to recognize the handwritten Nushu characters. The stroke relationships of a Nushu character reflect its structure, which can be statistically represented by the hidden markov model. Based on the prior knowledge of character structures, we design an adaptive statistical-structural character learning algorithm that accounts for the most important stroke relationships, which aims to improve the recognition rate by adapting selecting correct character to the current handwritten Nushu character condition. We penalize the structurally mismatched stroke relationships using the prior clique potentials and derive the likelihood clique potentials from Gaussian mixture models.

The rest paper is organized as follows. In section II, the proposed Nushu characters recognition scheme is proposed in detail. Detailed experimental results are shown in section III. Finally, the conclusion and future work are given in section IV.

II. Proposed Algorithm

A. Statistical-Structural Character Modeling

In the proposed modeling method, a character is represented by a set of model strokes. The structure of the model strokes is manually designed. As shown in Fig. 1, the model stroke is composed of a poly-line connecting K feature points. In Handwritten process, a model stroke is instantiated into various shapes of input strokes and, therefore, the feature points are instantiated into various pixels of the input strokes. In order to model such a variation, the feature point is represented by a distribution of the pixels. Because each pixel is identified by its position and direction, the feature point is represented by their distribution. Additionally, the direction at the feature point was also modeled to reflect more information.

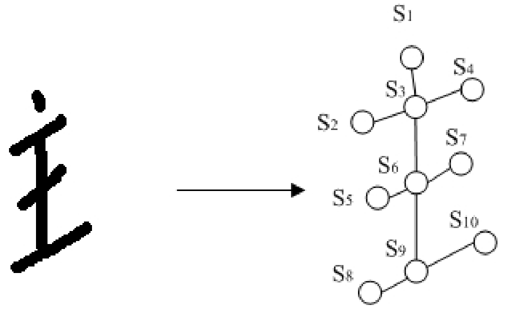


Figure 1. Statistical stroke model of Nushu character.

Denote the set of links of Nushu character Q, and let $|Q|$ denote the cardinality of the set Q. We define the stroke crosspoint as T_{event}. We refer to N consecutive crosspoints as n-graph. The special case of single crosspoints is referred to as unigraph. Two consecutive crosspoints are referred to as digraph in the literatures, and trigraph means three consecutive crosspoints, etc. Given a sequence of consecutive crosspoints $S=\{s_1,s_2,\cdots,s_m\}$, where m is the number of crosspoint sequence, we have n-graph with the size of $m-n+1$. We define the duration of n -graph $GD=\{d_1,d_2,\cdots,d_k \mid k\in N, k\in[1,m-n+1]\}$ as follows:

$$d_k = T_{event}^{s_{n+k-1}} - T_{event}^{s_k}. \tag{1}$$

The durations of n-graph are used as sequence features for further analysis in our proposed model. We make a natural assumption that the n-graph, with duration y, $P(y\mid q)$, forms a Gaussian distribution, such that:

$$P(y\mid q)=\frac{1}{\sqrt{2\pi}\sigma_q}e^{-\frac{(y-\mu_q)^2}{2\sigma_q^2}}, \tag{2}$$

where μ_q is the mean value of the duration y for n-graph , and σ_q is the standard deviation. Since behavioral characteristics of the individuals could be influenced by many reasons, the statistical analysis method used by previous work can be viewed as the same probability was given to the valid attempts of digraph latencies and durations within the standard deviations of the mean durations. By using Gaussian modeling, we can give higher probability to the n-graph durations of test samples that is more close to the n-graph mean durations of reference samples, and lower probability to the n-graph duration that is far from the mean of the n-graph for the reason that the individuals could be temporarily out of regular typing behavior, and we can take the irregular typing behavior without discarding the possibility that the set of n-graph durations provided by the corresponding individuals.

With the limitation that we are unable to collect all the typing crosspoints of the individual and calculate the exact parameters of the means and variances for each distinct combination of n-graph durations. We have to deduce $\{(\widehat{\mu_q},\widehat{\sigma_q})\}$ of n -graph durations, give a crosspoint sequence S, by the method of maximum likelihood estimation of the parameters. Fortunately, the maximum likelihood estimation of the parameters for Gaussian distribution can compute the sample mean and sample variance as follows.

$$\widehat{\mu_q}=\frac{\sum_{i=1}^{k}d_i(q)}{k}, \tag{3}$$

$$\widehat{\sigma_q}^2=\frac{\sum_{i=1}^{k}\left[d_i(q)-\widehat{\mu_q}\right]^2}{k-1}, \tag{4}$$

where is the number of n-graph q appeared in S.

In order to mine Nushu structural character, the proposed HMM models sequential data, such as the sequence of the crosspoints of Nushu characters and handwritten character information that we take into consideration. The HMM we use to model the structural character information of crosspoint and structural character sequence. HMMs are a modeling technique derived from Markov models, which are stochastic processes whose output is a sequence of states corresponding to some physical event. HMMs have the observation as a probabilistic function of the states, i.e. the resulting model is a doubly embedded stochastic process with an underlyings to chastic that is not observable (it is hidden), but can only be observed through another set of stochastic processes that produce these queue of observations.

Considering that it is a statistical graphical model, where each circle is a random variable. Unshaded circles q_t represent are unknown (hidden) state variables we wish to infer, and shaded circles y_t are observed state variables, where t is a specific point in time. A is a state transition matrix holding the probabilities of transitioning from q_t^i to q_{t+1}^j, where q^i means the i-th state. So we have:

$$P(q_{t+1}^j=1\mid q_t^i=1)=A_{ij}. \tag{5}$$

η is a state emission matrix holding the output probability $P(y_t\mid q_t^i=1)$ of i-th state. π_i is the initial state probability of i-th state. A compact notation $\lambda=(A,\eta,\pi)$ is used to indicate the complete parameter set of the model.

In our setting, given a crosspoint sequence S, n-graph G, $[n+1]$-graph G', such that:

$$S=\{s_1,s_2,\cdots,s_m\}, m\in N. \tag{6}$$

$$G=\{g_1,g_2,\cdots,g_{m-n+1}\} \tag{7}$$

$$G'=\{g_1^{'},g_2^{'},\cdots,g_{m-n}^{'}\} \tag{8}$$

The state transition matrix A is the probability of the frequency that the [$n+1$]-graph appeared in the as follows:

$$A_{g_t,g_{t+1}} = |g_t^{'}|/(m-n). \tag{9}$$

The state emission matrix η here is defined as the Gaussian distribution probability of the n -graph $G=\{g_1,g_2,\cdots,g_{m-n+1}\}$ with duration $GD=\{d_1(g_1),d_2(g_2),\cdots,d_{m-n+1}(g_{m-n+1})\}$ as follow:

$$\eta_g(d(g')) = \begin{cases} P(d(g')\,|\,g) = \dfrac{1}{\sqrt{2\pi}\sigma_g} e^{-\frac{(d(g')-\mu_g)^2}{2\sigma_g^2}} & ,g=g' \\ 0, otherwise \end{cases} \tag{10}$$

There are three basic problems to solve with the HMM $\lambda=(A,\eta,\pi)$:

1). Given a model parameters $\lambda=(A,\eta,\pi)$ and observation output sequence $O=O_1O_2\cdots O_t$, compute the probability $P(O\,|\,\lambda)$ of the observation output sequence.

2) Given a model parameters $\lambda=(A,\eta,\pi)$ and observation output sequence $O=O_1O_2\cdots O_t$, find the most probable state sequence $Q=Q_1Q_2\cdots Q_t$ which could have generated the observation output sequence.

3) Given an observation output sequence $O=O_1O_2\cdots O_t$, generate a HMM $\lambda=(A,\eta,\pi)$ to maximize the $P(O\,|\,\lambda)$.

We make the assumption that each individual has his/her own HMM with $\lambda=(A,\eta,\pi)$ for characters crosspoint and character structural characteristics. The problem to solve is that, given a crosspoint sequence and its character structural characteristics information, we have to choose one from the number of HMMs which has the highest probability to generate the crosspoint sequence S . Consequently, first we have to calculate the probability of crosspoint sequence S for each HMM. This is similar to the first basic problem to solve with HMM as described above, and we will show how to solve the problem with Forward algorithm.

The state probabilities α 's of each state can be computed by first calculating α for all states at $t=1$:

$$\alpha_1(g_1) = \pi(g_1)\cdot\eta_{g_1}(d_1). \tag{11}$$

Then for each time step $t=2,\cdots,k$, the state probability α is calculated recursively for each state:

$$\alpha_{t+1}(g_{t+1}) = \alpha_t(g_t)\cdot A_{g_t,g_{t+1}}\cdot\eta_{g_{t+1}}(d_{t+1}). \tag{12}$$

Finally, the probability of crosspoint sequence S given a HMM $\lambda=(A,\eta,\pi)$ is as follows:

$$P(S,G,GD\,|\,\lambda) = \alpha_k(g_k) = \alpha_{k-1}(g_{k-1})\cdot A_{g_{k-1},g_k}\cdot\eta_{g_k}(d_k). \tag{13}$$

The emission probabilities take less computation to obtain since we use the Gaussian distribution to model observed states. Additionally the observed states are only connected to the corresponding unknown states because we know the exact combination of n -graph the individual typed. So the summation of all partial probability of the state at time is ignored and only one probability is calculated.

In original version of the Forward algorithm, the computation involved in the calculation of $\alpha_t(j)$, $t\in[1,T], j\in[1,N]$, where T is the number of observations in the sequence and is the number of states in the model, requires $O(N^2T)$ calculations. In our modified version of the Forward algorithm, we can see that it only requires $O(NT)$ calculations.

B. Structural Learning

In the character building and extraction module, first we have to build the reference character for each Nushu character. It requires the user to provide the reference samples or handwritten profile. The more quantity of reference samples or histories provided, the more exact parameters can be extracted. After collecting sufficient number of reference samples, we use the maximum likelihood estimation for Gaussian modeling to calculate the parameters of each n -graph duration. We also have to compute the transition probability matrix and initial probability vector with respect to HMM. Then the parameters calculated for HMM are treated as the base element of the reference profile for each user. The feature building and extraction module extracts two observation sequences based on a sliding-window approach. One observation sequence is extracted in the horizontal direction, representing column observations, and the other one is extracted in the vertical direction, representing row observations. Each discrete observation represents a multidimensional feature vector, which is mapped by means of vector quantization (VQ).

The multidimensional feature vector combines both foreground and background information. The foreground features represents local information about the writing, observed from background–foreground transitions. The other two features represent a global point of view about the writing in the frame from which they are extracted. The background features are based on a configuration chain code, representing concavity information.

The learning algorithm includes three steps: setting up Nushu prototypes, initializing Nushu parameters, and the HMM parameter estimation. First, we set up Nushu prototypes for each category of characters using the observation from a well-segmented standard character, where the number of sites I of standard characters equals the number of labels J of the Nushu for each category.

The proposed adaptive learning algorithm maintains a control probability vector to select an accurate character among a set of characters at time. A good policy to update the probability vector is a pursuit algorithm that always rewards the action with the current minimum penalty estimate and that stochastic learning control performs well in speed of convergence. In this system, the probability vector is the rate selection probability vector $p(n) = [p_1(n),..., p_K(n)]$, where n is the index of the sequence of structural characters. The error of the nth handwritten character during is expressed by $\gamma(n)$. The set of characters available are $\{R_i : i = 1, 2, ..., K\}$. At beginning the $p(n)$ are assigned equal values:

$$p(0) = [1/K,...,1/K]. \qquad (14)$$

In order to maximize the likelihood, the recognition algorithm is required to find the index of the best character. Such an approach requires the knowledge of handwritten state during each recognition. The stochastic learning algorithm presented in this paper randomly selects a character. The character selection probability vector is altered by an iterative updating process, which maximizes the probability of maximizing the character recognition rate. Then, the character recognition proceeds with the fixed $p(n)$ until every character is selected at least M number of times after which $p(n)$ is augmented at each n. Following each recognition period, an update of $S(n)'$ and $p(n)$ are carried out considering the last M recognition signals of each recognition period. Then we can get:

$$S(n)' = \frac{R_i}{M} \sum_{j=L_i(n)-M+1}^{L_i(n)} I_i(j) \qquad (15)$$

where $I_i(j)$ is an indicator function:

$$I_i(j) = \begin{cases} 1, & if \ \ Recognization \ \ is \ \ correct \\ 0, & else \end{cases} \qquad (16)$$

$L_i(n)$ is the number of recognition periods for which the character structural R_i is selected during the nth recognition period.

The structural learning algorithm can be summarized as follows:

Step 1. If it is the first recognition period, initialize the probability vector as in equation (14). Else selects a character structural R_i ($i \in [1, K]$) according to probability distribution $p_i(n)$.

Step 2. Update $I_i(j)$ and $L_i(n)$. Then update $S(n)'$ according to (15).

Step 3. If for all $L_i(n) \geq M$ for all i go to next step, else go to step 1.

Step 4. Detect the index m' of the estimated best character structure and update according to the following equations:

$$p_i(n+1) = \begin{cases} p_i(n) - \Delta p, & i \neq m' \\ 1 - \sum\limits_{j=1, j \neq m'}^{K} p_i(n+1), & i = m' \end{cases} \qquad (17)$$

where Δp is a tunable penalty probability parameter.

The structural learning algorithm finds the index m' of the estimated best character $R_{m'}(n)$ maximizing the likelihood $S_{m'}(n)$ at time n:

$$m' = \arg\max_i \{R_i \sum_{k=L_i(n)-M+1}^{L_i(n)} I_i(k)\} \qquad (18)$$

For the probability of making the right decision, let the best recognition rate at time n, $R_m(n)$ be unique. Let $\phi_m(n)$ be the probability that the estimated best recognition rate is the actual best rate. Then we can get:

$$\phi_m(n) = \Pr\{ \sum_{k=L_i(n)-M+1}^{L_i(n)} I_i(k) < \frac{R_m(n)}{R_i} \sum_{k=L_m(n)-M+1}^{L_m(n)} I_m(k) \forall i \neq m(n)\} \qquad (19)$$

The above probability is readily obtained by using binomial probability distribution.

Since $M \geq \sum\limits_{k=L_i(n)-M+1}^{L_i(n)} I_i(k) \geq 0$ for all $i \in [1, K]$, when $\sum\limits_{k=L_i(n)-M+1}^{L_i(n)} I_i(k) > \sum\limits_{k=L_m(n)-M+1}^{L_m(n)} I_m(k)$, $\frac{R_m(n)}{R_i} \sum\limits_{k=L_m(n)-M+1}^{L_m(n)} I_m(k)$ may exceed M. Let's take into account the fact that $\sum\limits_{k=L_i(n)-M+1}^{L_i(n)} I_i(k) < \frac{R_m(n)}{R_i} \sum\limits_{k=L_m(n)-M+1}^{L_m(n)} I_m(k)$ in such cases.

Let ε_i be the largest nonnegative integer less than $\alpha(R_m(n)/R_i)$, where α is a nonnegative integer. Define the indicator function $O(\cdot)$ which value is 1 when condition within parentheses is satisfied, else is 0. Define the parameter θ_i for $i \in [1, K]$:

$$\theta_i = \varepsilon_i \cdot O(\varepsilon_i \leq M) + M \cdot O(\varepsilon_i > M) \qquad (20)$$

Then we have:

$$\phi_m(n) = \sum_{\alpha=1}^{M} [\Pr\{ \sum_{k=L_m(n)-M+1}^{L_m(n)} I_m(k) = \alpha\} \prod_{i=1, i \neq m}^{K} \sum_{\beta=0}^{\theta_i} \Pr\{ \sum_{k=L_i(n)-M+1}^{L_i(n)} I_i(k) = \alpha\}] \qquad (21)$$

Considering that:

$$\theta_i = M \ , \ \varepsilon_i \geq M \qquad (22)$$

We can get:

$$\phi_m(n) = \sum_{\alpha=1}^{M} \binom{M}{\alpha} Q_m^{\alpha} (1 - Q_m)^{M-\alpha} \cdot \prod_{i \neq m} \sum_{\beta=0}^{\theta_i} \binom{M}{\beta} Q_i^{\beta} (1 - Q_i)^{M-\beta} \qquad (23)$$

where Q_i is the probability of successful recognition of a Nushu character using the structural R_i.

Since the main objective in the work was to speed up the learning process, the proposed structural learning algorithm works by dividing an off-line training database into smaller blocks. Each iteration of the algorithm processes a different block of data. Thus, given an initial HMM, and the block data drawn from the training set, this algorithm works according to the algorithm.

HMMs are able to perform recognition tasks in pattern recognition systems. The most popular approach for such tasks consists of creating a set of HMMs so that each class is represented by an independent HMM. The classification of an unknown observation sequence $S=\{s_1,s_2,\cdots,s_m\}$, into a class, can be carried out by computing which HMM outputs the highest likelihood related to O. In detail, consider a class problem in which each class is represented by a single HMM. The likelihood can be easily computed by the forward–backward procedure.

C. Adjusting and Recogntion

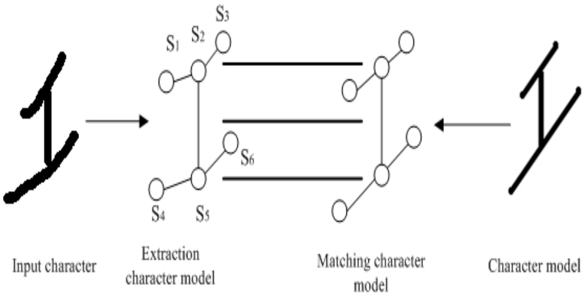


Figure 2. Nushu character adjusting and recognition.

In the adjusting module, given a crosspoint sequence S with claimed identity ID, we wish to examine the possibility that S generated by ID. First we transform the crosspoint sequence S to n-graph combinations G and calculate the character structural information of n-graph duration as usual. At this moment, we have $S=\{s_1,s_2,\cdots,s_m\}$, $G=\{g_1,g_2,\cdots,g_{m-n+1}\}$ and $GD=\{d_1,d_2,\cdots,d_{m-n+1}\}$. Now we produce a vector V, such that:

$$V=\{\mu_{g_1}-\varepsilon\sigma_{g_1},\mu_{g_2}-\varepsilon\sigma_{g_2},\cdots,\mu_{g_{m-n+1}}-\varepsilon\sigma_{g_{m-n+1}}\}, \quad (24)$$

where ε is the weighting factor, μ_{g_k} is ID's duration mean of n-graph g_k, and σ_{g_k} is ID's duration standard deviation of n-graph . V is the n-graph duration vector to evaluate the threshold value of the probability produced by the proposed modified forward algorithm. With the inputs GD, V, and λ_{ID}, we can apply the proposed forward algorithm mentioned above to obtain two probability value $P(S,G,GD\,|\,\lambda_{ID})$ and $P(S,G,V\,|\,\lambda_{ID})$. $P(S,G,GD\,|\,\lambda_{ID})$ can be viewed as the possibility if all the n-graphs durations in G are deviating ε times of duration σ from duration μ. $P(S,G,V\,|\,\lambda_{ID})$ is the threshold value of probability used to decide that the acceptance of the crosspoint sequence S is confirmed if following expression is true.

$$P(S,G,GD\,|\,\lambda_{ID})\geq P(S,G,V\,|\,\lambda_{ID}). \quad (25)$$

The weighting factor ε can be specified with respect to different level of security strength. In the Identification procedure, given a crosspoint sequence $S=\{s_1,s_2,\cdots,s_m\}$ from the individual and a set of HMMs $\lambda's=\{\lambda_1,\lambda_1,\cdots,\lambda_l\}$, where l is the number of HMM. The problem is to choose the best one from λ's which most probably generated S or there is no such one existed. In the beginning, the crosspoint sequence is transformed to n-graph combinations $G=\{g_1,g_2,\cdots,g_{m-n+1}\}$ and the timing information of n-graph duration $GD=\{d_1,d_2,\cdots,d_{m-n+1}\}$ is calculated. $P(S,G,GD\,|\,\lambda_{ID})$ for each HMM in λ's is produced by the proposed forward algorithm. We select user U with the maximum probability over others', such as:

$$P(S,G,GD\,|\,\lambda_U)=\max(P(S,G,GD\,|\,\lambda_j)),\ j\in[1,l]. \quad (26)$$

After that, we produce a vector for user U, such that V

$$V^U=\{\mu^U_{g_1}-\varepsilon\sigma^U_{g_1},\mu^U_{g_2}-\varepsilon\sigma^U_{g_2},\cdots,\mu^U_{g_{m-n+1}}-\varepsilon\sigma^U_{g_{m-n+1}}\}, \quad (27)$$

where ε is the weighting factor, $\mu^U_{g_1}$ is U's duration mean of n-graph g_k, and $\sigma^U_{g_k}$ is U's duration standard deviation of n-graph . Again we apply the proposed forward algorithm mentioned above to obtain two probability value $P(S,G,GD^U\,|\,\lambda_U)$ and $P(S,G,V^U\,|\,\lambda_U)$. If the expression $P(S,G,GD^U\,|\,\lambda_U)\geq P(S,G,V^U\,|\,\lambda_U)$, the crosspoint sequence generated by user U is confirmed. Otherwise, we consider the crosspoint sequence is not generated by any user in the user profile database.

III. Experimental Results

We evaluated the proposed algorithm on the Nushu database, which has 1783 classes with 200 samples for each class. Fig. 3 shows some typical samples in the Nushu database.

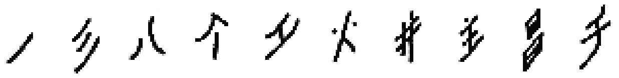


Figure 3. Samples in Nushu database.

We used a Matlab implementation on a PC with 2.4 GHz CPU and 1GB of memory. The average time on preprocessing is 0.002 seconds, and the stroke extraction (0.03 seconds) and the learning algorithm (0.01 seconds) consume a total of 0.04 seconds in the connected neighborhood system per character image. Although the structural match with one character model is efficient, requiring less than a second in our implementation, practically, we have to repeat the structural match with all categories of character models, such as 783 categories in Nushu database, to recognize one input character image. When the number of categories increases, the total time cost to recognize one character image increases. Currently, there are two commonly adopted strategies to expedite the recognition process. The first simultaneously uses several computers to perform the structural match with all the character models in parallel. The second is the hierarchical classification system that uses a fast algorithm to select a few candidate character models and then performs the structural match between the input strokes and these models to determine the best one.

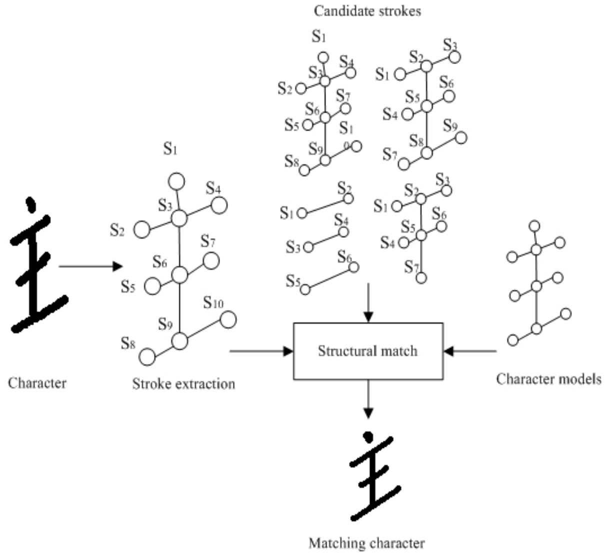


Figure 4. The stroke extraction and structural matching result of the proposed algorithm.

Fig. 4 shows the proposed stroke extraction and the structural matching results. The first column shows the input character. The second column shows the slant and moment normalization of the character skeleton. The third column shows the proposed character models, where the labels are numbered and the adjusting and recognition functions are performed. The structural learning algorithm assigns the best labels to the extracted candidate strokes.

We compared our method with the SCSM [3] and the attributed relational graph MBSEM [10]. The recognition rate of different scheme is shown in Fig. 5. The SCSM used the first 1000 odd number of samples of each category for training, and the first 2000 samples of even number of samples for test on Nushu database. By handling degraded region, the baseline recognition rate was 90.45 percent. For MBSEM, the recognition rate varies with the training samples increasing. The reason is that MBSEM can not recognize the handwritten Nushu characters, although training sample increases. For the proposed scheme, the recognition rate increases with the training samples increasing, which proves the proposed character structural learning algorithm can increase the recognition rate. It is clear that the recognition rate of the proposed scheme is 3.7% and 4.9% higher than those of SCSM and MBSEM respectively. The reason is that the proposed scheme not only takes the character structures but statistical-structural character into considerations, and the adaptive character structure learning algorithm guarantee the recognition rate. The HMM-based statistical-structural character modeling also truly depicts the Nushu character structure.

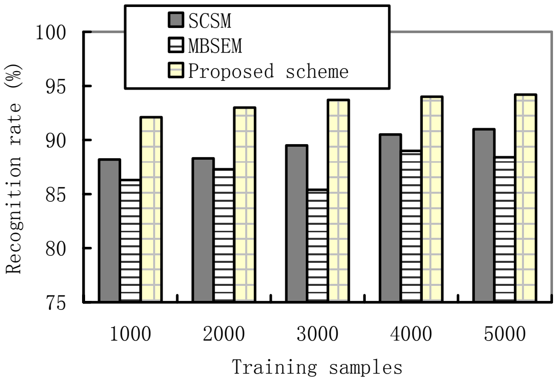


Figure 5. Recognition rate of different scheme.

IV. Conclusion and Future Work

A statistical-structural character learning algorithm based on hidden Markov model is proposed to recognize the handwritten Nushu characters. The approach is a convergence between statistical and structural threads of research. However, it improves the performance on both sides in term of overall recognition rate greatly. The stroke relationships of a Nushu character reflect its structure, which can be statistically represented by the hidden markov model. Based on the prior knowledge of character structures, an adaptive statistical-structural character learning algorithm accounts for the most important stroke relationships, which aims to improve the recognition rate by adapting selecting correct character to the current handwritten Nushu character condition. The experimental results and the comparisons with other methods show that the proposed method successfully detected and reflected the stroke relationships that seemed intuitively important. And the overall recognition rate is 93.7 percent, which is obviously higher than those of other schemes.

As a future research challenge, we will investigate how to decrease the use of the external knowledge for the proposed algorithm. For example, the use of a k-fold cross-validation would be useful to determine the number of iterations to train each block of data, Furthermore, topology learning could be employed to determine the best HMM topology.

ACKNOWLEDGMENT

This work is supported by the Natural Science Foundation of China (No. 60842004), the Natural Science Foundation of State Ethnic Affairs Commission (No. 08ZN02). The authors are grateful for the anonymous reviewers who made constructive comments.

REFERENCES

[1] Z.-B. GONG, "New Findings about Female Scripts," *Journal of South-Central University for Nationalities*, Vol. 23, No. 4, pp. 93-97, 2003.

[2] Z.-B. GONG, "Nushu in Jiangyong Is absolutely Not Ancient Characters during pre-Qin Day," *Journal of South-Central University for Nationalities*, Vol. 21, No. 6, pp. 130-133, 2001.

[3] I.-J. Kim, J.-H. Kim, "Statistical character structure modeling and its application to handwritten Chinese character recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, Is. 11, pp. 1422-1436, 2003.

[4] R. M. Suresh, S. Arumugam, "Fuzzy technique based recognition of handwritten characters," Image and Vision Computing, v 25, n 2, pp. 230-239, 2007.

[5] R. Zhang, X. Ding, H. L. Liu, 'Discriminative training based quadratic classifier for handwritten character recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, v 21, n 6, pp. 1035-1046, 2007.

[6] M. F. Zafar, O. Dzulkifli, "Writer independent online handwritten character recognition using a simple approach," *Information Technology Journal*, v 5, n 3, pp. 476-484, 2006.

[7] W. W. Lin, "Recognition of handwritten Chinese characters by feature matching," in Proc. 1991 Int. Conf. Computer Processing of Chinese and Oriental Languages, 1991, pp. 154–157.

[8] Y. L. Wu, T. M. Wu, and B. S. Jeng, "Optical Chinese character recognition using a projection profile and the Fourier transformation," J. Telecommun. Lab. Technique, vol. 20, pp. 137–145, 1990.

[9] H. D. Chang and J. F. Wang, "Preclassification for handwritten Chinese character recognition by a peripheral shape coding method," Pattern Recognition., vol. 26, pp. 711–719, 1993.

[10] C.L. Liu, I.J. Kim, and J.H. Kim, "Model-Based Stroke Extraction and Matching by Heuristic Search for Handwritten Chinese Character Recognition," Proc. Sixth Int'l Workshop Frontiers in Handwritten Recognition, pp. 547-556, 1998.

[11] X. Zhang and Y. Xia, "The Automatic Recognition of Handprinted Chinese Characters—A Method of Extracting an Order Sequence of Strokes," Pattern Recognition Letters, vol. 1, no. 4, pp. 259-265, 1983.

[12] H.Y. Kim and J.H. Kim, "Hierarchical Random Graph Representation of Handwritten Characters and Its Application to Hangul Recognition," Pattern Recognition, vol. 34, no. 2, pp. 187-201, 2001.

[13] I.-J. Kim and J.-H. Kim, "Statistical Character Structure Modeling and Its Application to Handwritten Chinese Character Recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 25, no. 11, pp. 1422-1436, Nov. 2003.

[14] M. Aksela, J. Laaksonen, "Adaptive combination of adaptive classifiers for handwritten character recognition," *Pattern Recognition Letters*, v 28, n 1, pp. 136-143, 2007.

[15] P. M. Patil, T. R. Sontakke, "Rotation, scale and translation invariant handwritten Devanagari numeral character recognition using general fuzzy neural network," *Pattern Recognition*, v 40, n 7, pp. 2110-2117, 2007.

Jiangqing Wang received the B.S. and M.S. degrees in Artificial Intelligence from Wuhan University, China, in 1986 and 1986, respectively; and Ph.D. degree in intelligent computation from Wuhan University, China, in 2007. She was a visiting professor of University of Wisconsin-La Crosse and Chonbuk National University. She is currently a Professor in College of Computer Science of South-Central University for Nationalities.

She has published over 40 papers in international journals and conferences in the areas of artificial intelligence and intelligent computation. Her current research interests are in the areas of character recognition, intelligent computation and optimization. The research activities have been supported by the Natural Science Foundation of China, Natural Science Foundation of State Ethnic Affairs Commission and Natural Science Foundation of Hubei province.

Dr Wang has been actively involved in around 20 international conferences, serving as Session Chair and a reviewer for numerous referred journals and many international conferences.

Rongbo Zhu received the B.S. and M.S. degrees in Electronic and Information Engineering from Wuhan University of Technology, China, in 2000 and 2003, respectively; and Ph.D. degree in communication and information systems from Shanghai Jiao Tong University, China, in 2006. He is currently an Associate Professor in College of Computer Science of South-Central University for Nationalities.

He has published over 40 papers in international journals and conferences in the areas of wireless communications, covering 3G mobile systems and beyond, MAC and routing protocols, and wireless ad hoc, sensor, and mesh networks. He received the Outstanding B. S. Thesis and M. S. Thesis awards from Wuhan University of Technology in 2000 and 2003, respectively. His current research interests are in the areas of wireless communications, protocol design and optimization. The research activities have been supported by the Natural Science Foundation of Hubei province and Natural Science Foundation of South-Central University for Nationalities.

Dr Zhu has been actively involved in around 10 international conferences, serving as Session Chair of the Intelligent Networks Track at LSMS'07, and as a reviewer for numerous referred journals such as IEEE Communication Letters, Wiley Wireless Communications and Mobile Computing, and many international conferences such as IEEE Globecom'08, IEEE ICC'07, IET CCWMSN'07, ISICA'07 and so on.

Semi-supervised Learning for SVM-KNN

Kunlun Li
College of Electronics and information Engineering, Hebei University, Baoding, 071002,china
Email: likunlun@hbu.edu.cn

Xuerong Luo and Ming Jin
College of Electronics and information Engineering, Hebei University, Baoding, 071002,china
School of Telecommunication Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876,china
Email: {baoding0312@163.com, jinan527@yahoo.com.cn}

***Abstract*—Compared with labeled data, unlabeled data are significantly easier to obtain. Currently, classification of unlabeled data is an open issue. In this paper a novel SVM-KNN classification methodology based on Semi-supervised learning is proposed, we consider the problem of using a large number of unlabeled data to boost performance of the classifier when only a small set of labeled examples is available. We use the few labeled data to train a weaker SVM classifier and make use of the boundary vectors to improve the weaker SVM iteratively by introducing KNN. Using KNN classifier doesn't enlarge the number of training examples only, but also improves the quality of the new training examples which are transformed from the boundary vectors. Experiments on UCI data sets show that the proposed methodology can evidently improve the accuracy of the final SVM classifier by tuning the parameters and can reduce the cost of labeling unlabeled examples.**

***Index Terms*—semi-supervised learning, support vector machine, K-nearest neighbor, boundary vectors**

I. Introduction

In this paper we focus on solving the classification problem by using semi-supervised learning strategy. Traditional classifiers are constructed based on labeled data in supervised learning. Labeled examples, however, are often difficult, expensive, or time consuming to obtain, as they require the efforts of experienced human annotators. Meanwhile it is relatively easier to collect unlabeled examples and there have been a few classification approaches using unlabeled data in recent years. Semi-supervised learning addresses this problem by using large number of unlabeled data, together with the small number of labeled data, to construct better classifiers. Because semi-supervised learning requires less human effort and gives higher accuracy, it is of great interest both in theory and in practice [1].

Traditional supervised learning needs sufficient labeled data as training sets, or else can't get a supervised learning method with strong generalization [2], but obtaining lots of labeled data is difficult in practice, even can't come true. Unsupervised learning tries to find the inner structure of the unlabeled data to construct the corresponding learning machine, so it leads to unsupervised learning can't ensure high learning accuracy usually [3]. In this case, only using traditional machine learning strategy can't gain a learning machine with strong generalization and high accuracy if there is inadequate labeled data.

Semi-supervised learning is brought forward as a learning strategy in recent years, which not only makes use of the labeled data and unlabeled data but also supplements the shortages of supervised learning and unsupervised learning. Semi-supervised learning theory and algorithm developed quickly in the recent years [4], because it has been become research focus in the field of machine learning, attracting much more scholars devote themselves to the further study.

The central issue that this paper addresses is how to use information from unlabeled data to enhance the predictability of classification. In this paper, we propose a novel SVM-KNN classification method based on semi-supervised learning, which makes full use of unlabeled data. Our experimental results support the statistical learning theory showing that incorporating unlabeled data improves accuracy of the classifier when insufficient training information is available.

This paper is organized as follows. In section II we describe the related works for our method. In Section Ⅲ we introduce the proposed semi-supervised learning methodology. In section Ⅳ we present some experimental results, using a tuning method that utilizes both labeled and unlabeled data to enhance the accuracy of classification. We gieve conclusion in section Ⅴ.

II. Related work

A. Semi-supervised learning

Semi-supervised learning (SSL) is halfway between supervised and unsupervised learning [3]. In addition to

This work is supported by the National Natural Science Foundation of China (No.60773062 & No. 60873100), the Science and Technology Supporting Program of Hebei Province (No. 072135188) and Scientific research plan projects of Hebei Educational Bureau (No. 2008312).

Manuscript received January 1, 2008; revised June 1, 2008; accepted July 1, 2008. The corresponding author is Kunlun Li.

doi:10.4304/jcp.5.5.671-678

unlabeled data, this kind of methodology is provided with some supervised information – but not necessarily for all examples. Often, this information will be the labels associated with some of the examples. In this case, the data of SSL set $X=(x_i)_{i\in[n]}$ can be separated into two parts: the points $X_h=(x_1,\ldots,x_h)$, for which labels $Y_h=(y_1,\ldots,y_h)$ are provided, and the points $X_t=(x_{h+1},\ldots,x_{h+t})$, the labels of which are not known. This is the normal form of semi-supervised learning set [4].

Semi-supervised learning will be most useful whenever there are much more unlabeled data than labeled. This is likely to occur if obtaining data points is cheap, but obtaining the labels costs a lot of time, effort, or money [5]. This is the instance in many application areas of machine learning, for example, in speech recognition, it costs almost nothing to record large amounts of speech, but labeling it requires many people to listen to it and type a transcript. Since unlabeled data contain less information than labeled data, they are required in large amounts in order to increase prediction accuracy significantly [4].

A number of classification algorithms that uses both labeled and unlabeled data have been proposed, for example, self-learning or self-labeling is the earliest semi-supervised learning method Probably, which is still extensively used in the processing of natural language. S ^{3} VM, originally called Transductive SVM, they are now called Semi-Supervised SVM to emphasize the fact that they are not capable of transduction only, but also can induction. The idea is to find a decision boundary in 'low density' regions. Graph-based algorithms, one can build a weighted graph over the labeled and unlabeled examples, and assume that two strongly-connected examples tend to have the same label and solve an optimization problem. Generative models, Mixture of Gaussian or multinomial distributions, and pretty much any generative model can do semi-supervised learning. Especially for EM algorithm, which is often used for training generative models when there is unlabeled data [1].

One of the many approaches to semi-supervised learning is to first train a weaker predictor, which is then used in exploiting the unlabeled examples. For instance, in content-based image retrieval (CBIR), a user usually poses several example images as a query and asks a system to return similar images. In this situation there are many unlabeled examples, i.e. images that exist in a database, but there are only several labeled examples, i.e. the query images. Another instance is online web page recommendation. When a user is surfing the Internet, he may occasionally encounter some interesting web pages and may want the system bring him to similarly interesting web pages. It will be difficult to require the user to confirm more interesting pages as training examples because the user may not know where they are. In this instance, although there are a lot of unlabeled examples, i.e. web pages on the Internet, there are only a few labeled examples, i.e. the current interesting web pages [6]. In these situations, there are very few labeled training examples to rely on. The initial weaker predictor may not classify the other examples correctly, so we propose a novel SVM-KNN classification method based on semi-supervised learning to solve these cases.

B. Support vector machine

Support vector machine (SVM) is the youngest part in the statistical learning theory [7], whose dominating content is accomplished from 1992 to 1995 and developed quickly at present because of its solid theory and widespread applications. SVM is based on the structural risk minimization principle (SRM), which was proposed by Vapnik in 1998. Comparing with other learning methods, its generalization is optimal.

SVM is proposed through the optimal hyperplane in the linear partition case [8], the optimal hyperplane is depicted in the figure1, the pentacles and squares denote the training examples of two classes respectively, L is the separated line which partitions two classes correctly, L_1 and L_2 are the separate lines nearest to and parallel with L, the distance between L_1 and L_2 is called classification margin. The optimal classification line requests to partition two classes correctly and make the margin up to maximum so as to ensure structural risk minimization [9]. Extend to high dimension feature space, the optimal line becomes the optimal hyperplane.

The basic classification task is to estimate a classification function $f: R^N \to \{\pm 1\}$ using input-output training examples [10] from two classes

$$(x_{i,}y_i)\ \ i=1,\ldots,\ \mathrm{n}, \mathrm{x}\in R^d, y\in\{+1,-1\} \qquad (1)$$

The function f should correctly classify unseen examples (x,y) i.e. $f(x)=y$ if (x,y) have the same probability distribution with the training data. In this work we will discuss binary classification [10]. If the points are linearly separable, then there exist an n-vector w and scalar b such that

$$y_i\left[\left(\omega\cdot x_i\right)+b\right]-1\geq 0,\ i=1,2,\ldots,\ \mathrm{n} \qquad (2)$$

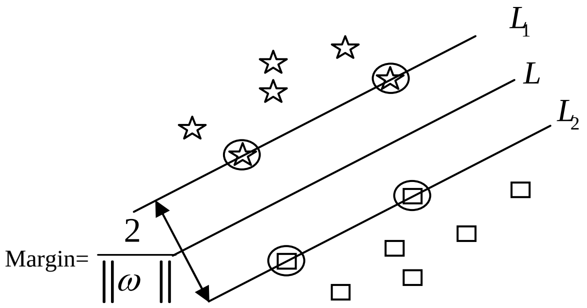


Figure1.The optimal hyperplane

The "optimal" separating plane, $\omega \cdot x+b=0$, is the one which is furthest from the closest points in the two classes [11]. Geometrically this is equivalent to maximizing the separation margin or distance between the two parallel planes $w \cdot x + b = \pm 1$ i.e. L_1, L_2 (see Figure 1.)

The "margin of separation" in Euclidean distance is $2/\|\omega\|_2$ where $\|w\|_2 = \sum_{i=1}^{n} w_i^2$ is the 2-norm. To maximize the margin, we minimize $\|w\|$ subject to the constraints (2). According to structural risk minimization[12], for a fixed empirical misclassification rate, larger margins should lead to better generalization and prevent over-fitting in high-dimensional feature spaces, so the task of standard SVM is:

$$\min \Phi(\omega) = \frac{1}{2}\|\omega\|^2 = \frac{1}{2}(\omega \cdot \omega)$$
$$s.t. y_i \left[\left(\omega \cdot x_i \right) + b \right] - 1 \geq 0,\ i = 1, 2, \ldots,\ \text{n} \quad (3)$$

The Lagrangian equation is:

$$\begin{aligned} L(w,b,\alpha) &= \frac{1}{2}(w \cdot w) \\ &- \sum_{i=1}^{n} \alpha_i \{ y_i [(w \cdot x_i) + b] - 1 \} \end{aligned} \quad (4)$$

Then the former of the optimal problem accordingly becomes to:

$$\begin{cases} \min L(\alpha) = 1/2 \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N} \alpha_i \\ s.t \quad \alpha_i \geq 0, \quad i = 1,2,\ldots,N \\ \sum_{i=1}^{N} y_i \alpha_i = 0 \end{cases} \quad (5)$$

At last it calculates the decision function[13]:

$$\begin{aligned} f(x) &= \operatorname{sgn}\{ (w^* \cdot x) + b^* \} \\ &= \operatorname{sgn}\{ \sum_{i=1}^{n} \alpha_i^* y_i (x_i \cdot x) + b^* \} \end{aligned} \quad (6)$$

The above only considers the linearly separable case. But in practice the most data is nonlinearly separable. In order to solve the nonlinearly separable cases, we introduce the kernel function into SVM. Generally to say, through the space mapping of the data, the dimensional-low data will be mapped into a sufficiently high dimensional space. Then the data can be linearly separable in the high dimensional space. To avoid the complex inner product operation of the high dimensional space, the kernel function uses the simple operation of the original space to replace it [14] [15]. Then:

$$K(x, x_i) = \varphi(x) \cdot \varphi(x_i) \quad (7)$$

The quadratic programming of the classical SVM is:

$$\begin{cases} \min_{\alpha_i} Q(\alpha_i; \varphi(x_i)) = \frac{1}{2} \sum_{i,j=1}^{N} y_i y_l \varphi(x_i)^T \varphi(x_j) \alpha_i \alpha_j - \sum_{i=1}^{N} \alpha_i \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq c, \qquad i = 1, 2, \cdots, N \end{cases} \quad (8)$$

The final classifier is:

$$f(x) = \operatorname{sgn}\left[\sum_{i=1}^{N} \alpha_i y_i K(x \cdot x_i) + b \right] \quad (9)$$

C. Selection of Kernel function

There are four common kernel functions depicted as below, we must decide which one to try first. Then the penalty parameter C and kernel parameters are chosen.

- Linear: $K(x_i, x_j) = x_i^T x_j$.
- Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$.
- Radial basis function (RBF):
  $$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0$$
- Sigmoid: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$.

In our experiment, we choose RBF kernel as our kernel function. The RBF kernel nonlinearly maps examples into a higher dimensional space, unlike the linear kernel, it can handle the situation when the relation between class labels and attributes is nonlinear. What is more, the linear kernel is a special case of RBF as [16] [17] shows that the linear kernel with a penalty parameter has the same performance as the RBF kernel with some parameters. In addition, the sigmoid kernel behaves like RBF for certain parameters [17] and the number of hyper-parameters which influences the complexity of model selection. The polynomial kernel has more hyperparameters than the RBF kernel [18].

D. K-Nearest neighbor

The K-Nearest Neighbor (KNN) algorithm is proposed by Cover and Hart in 1968 [19], whose theory has been developed maturely. K nearest neighbors are calculated using Euclidean distance, though other measures are available, Euclidean distance offers a fine mix of ease and efficiency. The classification of the example is determined by a majority vote of the labels of the *k*-near neighbors [20]. Intuitively, This method is very simple: for instance, if example x_1 has k nearest examples in the feature space and a majority of them have the same label y_1, then example x_1 belongs to y_1.

Although KNN method depends on utmost theorem in the theory, during the decision course it is only related to small number of nearest neighbors, so adopting this method can avoid the problem of examples imbalanced, otherwise, KNN mainly depends on limited number of nearest neighbors around not a decision boundary, so it is suitable for classifying the case of examples set of boundary intercross and examples overlapped.

Euclidean distance is calculated as follows [21]: suppose two vectors x_i and x_j , $x_i = (x_i^1, x_i^2, \ldots, x_i^n)$, $x_j = (x_j^1, x_j^2, \ldots, x_j^n)$,

the distance between x_i and x_j is :

$$D(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_i^k - x_j^k)^2} \tag{10}$$

In our experiment, we estimate the nearest neighbor of an example according to this formula.

III. SVM-KNN METHODOLOGY BASED ON SEMI-SUPERVISED LEARNING

A. The motivation of the methodology

In many pattern classification problems, if there is plenty of unlabeled data while only a small number of labeled data is available, then we should adopt semi-supervised learning strategy, the existing methods have different kinds of restrictions at present, so how to seek an approach to solve the classification problem extensively is a troublesome problem. We design a semi-supervised learning methodology by combining SVM and KNN algorithm, SVM fails to provide an accurate estimate of the true decision boundaries, because of the small size of labeled data. In contrast, we can utilize the information from the unlabeled data, which can help to recover the true decision boundaries for classification. As for SVM classification, the support vectors decide the decision boundaries directly, while the boundary vectors stand a good chance to be the support vectors, so we can choose the boundary vectors to rectify the decision boundaries iteratively. We employ KNN algorithm to label the boundary vectors because KNN mainly depends on limited number of nearest neighbors around, so it is suitable for classifying the case of examples set of boundary intercross and examples overlapped. At last the boundary vectors are mingled with the initial training examples to improve the accuracy of classification.

The primary goal of this paper is to develop a semi-supervised learning methodology to show high performance of classification by utilizing unlabeled data. Unlike existing methods, our methodology is planed to adapt to a variety of cases unlike other approaches have many restrictions. It yields an improvement when unlabeled data can help to reconstruct the optimal classification boundary by tuning three parameters. The three parameters can be depicted in the section

B. Proposed methodology

When we classified a data set including large number of unlabeled data, if only utilize the few training examples available, then we can't obtain a high accuracy classifier with inadequate training examples; if we want to obtain a classifier of high performance, then labeling the unlabeled data is necessary, but labeling vast unlabeled data wastes time and consumes strength. In this paper, we propose a novel method which uses SVM cooperated with KNN for classification based on semi-supervised learning theory. The general model is depicted as above (See Figure 2). To begin with, we construct a weaker classifier SVM according to the few training examples available, then using the weaker SVM classifies the remaining large number of unlabeled data in the data set, picking out n examples belonging to each class around the decision boundary by calculating Euclidean distance in the feature space, because the examples located around the boundary are easy to be misclassified, but they are likely to be the support vectors, we call them boundary vectors, so picking out these boundary vectors whose labels are fuzzy labeled by the weaker classifier SVM. Secondly we recognize these boundary vectors as testing set while recognize initial training examples as training set, use KNN method to classify them and recognize the results as the labels for boundary vectors. In the end, we put these boundary vectors and their labels into initial training set to enlarge the number of the training examples then retrain a SVM, iteratively until the number of the training examples is m times of the whole data set. The experimental results on three UCI data sets indicate that the final classifier SVM has significant improvement on accuracy.

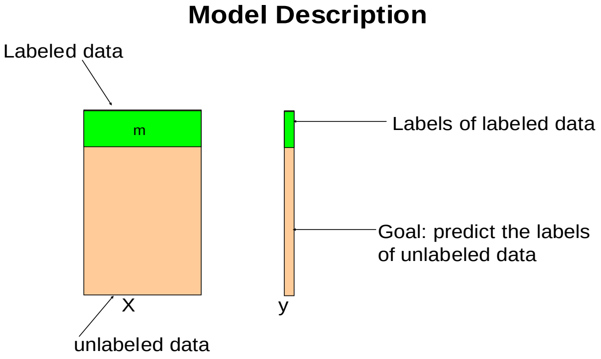


Figure2. The general description of the proposed model

The detailed steps of our method are as follows:

1) Utilize the labeled data available in a data set as initial training set and construct a weaker classifier SVM1 based on this training set.
2) Utilize SVM1 to predict the labels of all the remaining unlabeled data in the data set, then pick out 2 n examples located around the decision boundary as boundary vectors.
    a) Choose an example x_i from the class of A (A is the label) and calculate the distance between x_i and all the examples of class B (B is the label) using Euclidean distance subsequently pick out n examples of B corresponding to the n minimum distances.
    b) Choose an example y_i from the class of B (B is the label) and calculate the distance between y_i and all the examples of class A (A is the label) using Euclidean distance subsequently pick out n examples of A corresponding to

the n minimum distances.

c) We call the $2n$ examples as boundary vectors, make the $2n$ boundary vectors together as a new testing set.

3) KNN classifier classifies the new testing set with the initial training set, the boundary vectors get new labels.

4) Put the boundary vectors and their new labels into initial training set to enlarge the training set, then retrain a new SVM2.

5) Iteratively as above until the number of the training examples is m times of the whole data set.

The final SVM predicts the initial unlabeled data and the remaining unlabeled data, the results indicating that it has significant improvement.

Ⅳ . EXPERIMENTAL RESULTS

Our experiment was carried out on three publicly available labeled data sets. Unlabeled data was simulated by dropping labels from some points in a given data set. In a given data set, some examples are randomly picked out to be used as the labeled training examples, while the remaining data are used as the unlabeled examples. The procedure is repeated ten times with random data partitions and reports the average result.

The three benchmark data sets used in the following experiments are all chosen from the UCI machine learning repository [22], they are Iris data set, Breast cancer data set and Ionosphere data set. Iris data set: This data set consists of 150 four-dimensional examples. It is divided into three classes of equal size 50, but we only choose the examples of the two non-linear classes. The four features of this data set are: sepal length, sepal width, petal length, and petal width. Breast cancer data set: This is a nine-dimensional data set with 683 examples in two classes. There are 444 examples in class ‘‘benign’’, and 239 examples in class ‘‘malignant’’. The nine features are: clump thickness, cell size uniformity, cell shape uniformity, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses. Ionosphere data set: This is a 34-dimensional data set with 351 examples in two classes. All 34 attributes are continuous and the two classes are “good” and “bad”.

1） Experiment on Iris data

In the experiment, we choose all the examples of the two non-linear classes as the data set in advance, subsequently choose 10 labeled examples randomly from the data set as initial training set and get rid of the labels of the other 90 examples as unlabeled data set. We repeat the procedure ten times and report the average result.

Table 1 Experiment result on Iris data when m =0.26 n =2

k	Training examples	Testing examples	Accuracy on initial unlabeled data (%)
	10	90	91.11%
2	26	74	95.65%
4	26	74	94.20%
6	26	74	95.65%

Table 2 Experiment result on Iris data when k =1 n =2

m	Training examples	Testing examples	Accuracy on initial unlabeled data (%)
	10	90	91.11%
0.26	26	74	95.65%
0.34	36	64	96.88%
0.40	36	64	94.92%

Table 3 Experiment result on Iris data when m =0.4 k =1

n	Training examples	Testing examples	Accuracy on initial unlabeled data (%)
	10	90	91.11%
3	22	78	98.15%
5	40	60	93.10%
8	26	74	90.00%

It is shown in the three Tables that we choose 10% labeled data of the whole data set as initial training set, after iterations the final training examples become more at a different degree so that it improves the classification accuracy on the 90% unlabeled data, during the course of the experiment we can tune the three parameters to obtain the optimal result as Table 3 shows when m =0.4 k =1 n =3.

2） Experiment on Breast cancer data

In the experiment, we choose 100 data randomly from the data set as initial training set and get rid of the labels of the other 583 examples as unlabeled data set. We repeat the procedure ten times and report the average result.

Table 4 Experiment result on breast cancer data when m =0.3 n =3

k	Training examples	Testing examples	Accuracy on initial unlabeled data (%)
	100	583	78.22%
1	203	480	80.34%
5	203	480	80.77%
8	203	480	80.77%

Table 5 Experiment result on breast cancer data when k =5 n =3

m	Training examples	Testing examples	Accuracy on initial unlabeled data (%)
	100	583	78.22%
0.3	203	480	80.77%
0.4	272	411	80.69%
0.5	335	348	84.55%

Table 6 Experiment result on breast cancer data when m =0.5 k =5

n	Training examples	Testing examples	Accuracy on initial unlabeled data (%)
	100	583	78.22%
2	341	342	84.91%
5	331	352	90.06%
8	341	342	86.24%

It is shown in the three Tables that we choose 15% labeled data of the whole data set as initial training set, after iterations the final training examples become more at a different degree so that it improves the classification accuracy on the 85% unlabeled data, during the course of the experiment we can tune the three parameters to obtain the optimal result as Table 3 shows when m =0.5 k =5 n =5.

3） Experiment on Ionosphere data

In the experiment, we choose 70 data randomly from the dataset as initial training set, using the other 281data remained as unlabeled data. We repeat the procedure ten times and report the average result.

Table 7 Experiment result on Ionosphere data when m =0.3 n =3

k	Training examples	Testing examples	Accuracy on initial unlabeled data (%)
	70	281	69.40%
2	100	251	72.24%
3	100	251	71.72%
5	100	251	72.65%

Table 8 Experiment result on Ionosphere data when k =3 n =3

m	Training examples	Testing examples	Accuracy on initial unlabeled data (%)
	70	281	69.40%
0.3	100	251	72.65%
0.4	136	215	77.14%
0.45	152	199	77.84%

Table 9 Experiment result on Ionosphere data when m =0.45 k =3

n	Training examples	Testing examples	Accuracy on initial unlabeled data (%)
	70	281	69.40%
1	156	195	80.14%
2	154	197	79.27%
5	150	201	79.69%

It is shown in the three Tables that we choose 20% labeled data of the whole data set as initial training set, after iterations the final training examples become more at a different degree so that it improves the classification accuracy on the 80% unlabeled data, during the course of the experiment we can tune the three parameters to obtain the optimal result as Table 3 shows when m =0.45 k =3 n =1.

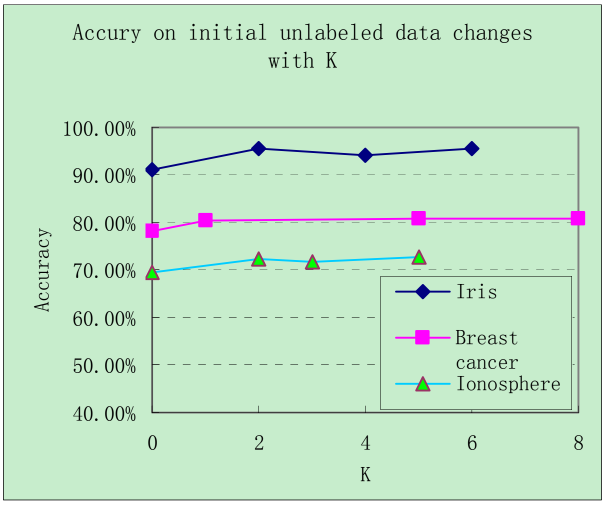


Figure3. When the other two parameters are fixed, the accuracy on initial unlabeled data changes along with the change of the K and we can choose the optimal value according to the above figure.

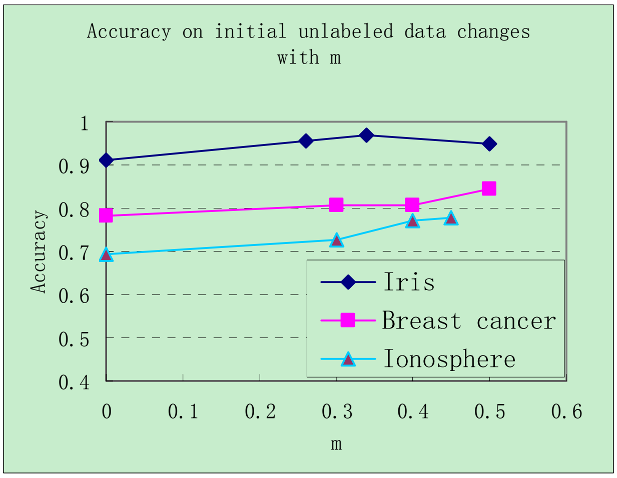


Figure4. When the other two parameters are fixed, the accuracy on initial unlabeled data changes along with the change of the m and we can choose the optimal value according to the above figure.

In all the experiments, the classes of unlabeled data are binary and the initial training set consists of the examples of two classes, during the course of experiment we set three parameters including k, m, n , k is the number of nearest neighbors, m is the percentage controlling the number of training data account for the number of whole data, n is the number of boundary vectors picked out from every class for each iteration. As depicted above we can tune the three parameters in order to gain better performance.

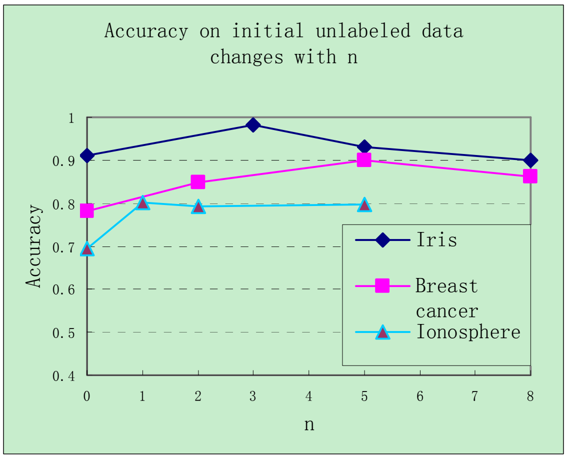


Figure5. When the other two parameters are fixed, the accuracy on initial unlabeled data changes along with the change of the n and we can choose the optimal value according to the above figure.

In a word, we propose a novel SVM-KNN classification methodology based on semi-supervised learning theory improve the accuracy of the final classifier. We only introduce part of the unlabeled-labeled data into the training set, but these unlabeled-labeled data are boundary vectors picked out from the classification boundary, the boundary vectors may be the support vectors so the final SVM classifier can predict the examples excellently.

Ⅴ. CONCLUSION

In this work we have described a classification method based on semi-supervised learning theory in which unlabeled data can be used to augment the training set and improve the accuracy of the final classifier. Our results support the statistical learning theory results that incorporating unlabeled data improves accuracy of the classifier when insufficient training information is available. In order to obtain better results the three parameters can be tuned conveniently according to practical environment. The preliminary experimental results presented suggest that this method of utilizing large number of unlabeled data has a potential for significant benefits in practice [6].

Many research questions remain. In the future we will study the classification problem with unlabeled data set in which the labels of unlabeled examples are unbalanced distributed and solve the multi-category unlabeled data classification, so the further work are clearly required.

ACKNOWLEDGMENTS

We would like to express our gratitude to all those who gave us help to complete this paper, especially for Mr Wei Zhang, Zheng Cao and Chao Zhang whose constructive suggestions and encouragements help us in all the time of research and writing of this paper.

REFERENCES

[1] X.J. Zhu. Semi-supervised learning literature survey[R]. Technical Report 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, December, 2007.

[2] Miller, D. J., and Uyar, H. S. A mixture of experts classifier with learning based on both labeled and unlabeled data. Advance in NIPS 9.571–577, 1997.

[3] Fung, G., & Mangasarian, O. Semi-supervised support vector machines for unlabeled data classification (Technical Report 99-05). Data Mining Institute, University of Wisconsin Madison, 1999.

[4] Olivier Chapelle, Bernhard Schölkopf, Alexander Zien. Semi- Supervised Learning [M]. The MIT Press, 2006

[5] Blum, A., and Mitchell, T. Combining labeled and unlabeled data with co-training. In COLT, 92–100, 1998.

[6] Zhou, Z.-H., Zhan, D.-C., & Yang, Q. Semi-supervised learning with very few labeled training examples. Twenty-Second AAAI Conference on Artificial Intelligence (AAAI-07), 2007.

[7] Bennett, K., & Demiriz, A. Semi-supervised support vector machines.Advances in Neural Information Processing Systems, 11, 368–374, 1999.

[8] Vapnik V. The Nature of Statistical Learning Theory [M]. New York: Springer-Verlag, 2000.

[9] P. S. Bradley and O. L. Mangasarian. Massive data discrimination via linear support vector machines. Technical Report Mathematical Programming Technical Report 98-05, University of Wisconsin-Madison, 1998.

[10] C. Burges, "A tutorial on support vector machines for pattern recognition", Data Mining and Knowledge Discovery, vol. 2, 1998.

[11] V. N. Vapnik, S. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation and signal processing", Advances in Neural Information Processing Systems, vol. 9, Cambridge, Mass: MIT Press, 1997.

[12] Y. Q. Zhang and D. G. Shen, "Design efficient support vector machine for fast classification", Pattern Recognition, vol. 38, 157-161, 2005.

[13] G. Fung, O.L. Mangasarian, Semi-supervised support vector machines for unlabeled data classification, Optim. Methods Software15 (1) (2001) 29–44.

[14] S.R. Waterhouse, A.J. Robinson, Classification using hierarchical mixtures of experts, in: Proceedings of the 1994 IEEE Workshop on Neural Networks for Signal Processing, pp. 177–186, 1994.

[15] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods [M]. Cambridge University Press, 2000.

[16] Keerthi, S. S. and C.-J. Lin (2003). Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Computation 15 (7), 1667{1689.

[17] Lin, H.-T. and C.-J. Lin (2003). A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. Technical report, Department of Computer Science, National Taiwan University.

[18] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification.Technical report, National Taiwan University,2003.

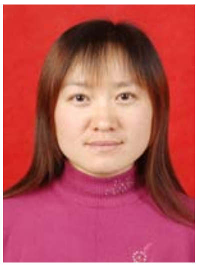
[19] Dasarathy, B. V., Nearest Neighbor (NN) Norms, NN Pattern Classification Techniques. IEEE Computer Society Press, 1990.

[20] Wettschereck, D., Dietterich, T. G. “An Experimental Comparison of the Nearest Neighbor and Nearest-hyperrectangle Algorithms,” Machine Learning, 9: 5-28, 1995.

[21] Platt J C. Fast Training of Support Vector Machines Using Sequential Minimal Optimization [M]. Advances in Kernel Methods:Support Vector Machines (Edited by Scholkopf B,Burges C,Smola A)[M]. Cambridge MA: MIT Press, 185-208, 1998.

[22] UCI repository of machine learning databases: http://www.ics.uci.edu/mlearn/MLRepository.

KunLun Li, born in 1962, received the PhD degrees in Signal & Information Processing from Beijing Jiaotong University, China, in 2004 and join College of Electronic and Information Engineering of Hebei University as the associate professor at present. His main research interests include machine learning, data mining, intelligent network security and biology information technology. In these areas, he has published over 20 technical papers in refereed international journals or conference proceedings.



Xuerong Luo born in 1982, received her B.Sc. degree in College of Physics Science & Technology, Hebei University, Baoding, China. Currently she is a M.Sc. candidate in College of Electronics and information Engineering, Hebei University, Baoding, 071002, china. Her main research interests include pattern recognition and artificial intelligent, machine learning and data mining, information security.

Ming Jin born in 1983, received his B.Sc. degree in College of Physics Science & Technology, Hebei University, Baoding, 071002, China. Currently he is a M.Sc. candidate in School of Telecommunication Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, china. His main research interests include mobile telecommunication and signal process especially about system level simulation in TD-SCDMA and HSDPA.

Exploration on Feature Extraction Schemes and Classifiers for Shaft Testing System

Kyungmi Lee
School of Business, James Cook University, Cairns, Queensland 4870, Australia
Email: Joanne.Lee@jcu.edu.au

***Abstract*— A-scans from ultrasonic testing of long shafts are complex signals, thus the discrimination of different types of echoes is of importance for non-destructive testing and equipment maintenance. Research has focused on selecting features of physical significance or exploring classifier like Artificial Neural Networks and Support Vector Machines. This paper summarizes and reports on our comprehensive exploration on efficient feature extraction schemes and classifiers for shaft testing system and further on the diverse possibilities of heterogeneous and homogeneous ensembles.**

***Index Terms*—Signal Classification, Non-Destructive Testing, Signal Feature Extraction**

I. Introduction

Applications of machine learning demand exploration of feature extraction methods and classifier types in order to obtain systems with reliable high accuracy. The industrial application discussed in this paper is the classification of ultrasonic echoes in an A-scan for testing shafts. The application is particularly challenging as A-scans are taken from the end of a long large complex shaft. Although several pattern analysis and machine learning techniques have been used with success in analyzing ultrasonic A-scan data [1-2], they are typically in the context of very short signals where the task is simply detecting the existence of an echo indicating a fault in the material. In long shafts there are many kinds of echoes including echoes where there is no fault. These *mode-converted* echoes are the result of reflection and other artifacts of the ultrasonic signal navigating and filling the shaft. They may cause misjudgement of the position of real faults (cracks) of shafts, thus it is important to distinguish them from genuine echoes.

Therefore, the problem is to discriminate efficiently the different types of reflectors among the large volumes of ultrasonic shaft test data, and classify them into a) those that correspond to flaws, cracks and other defects (CR) and b) the multiple reflections and mode-converted echoes (MC) of other reflectors. A main problem in the field is that the signal echoes caused by CR can be confused with fainted echoes caused by MC and vice versa. Consequences of misclassification are catastrophic with enormous cost in downtime, consequential damage to associate equipment and potential injury to personnel [3]. Conventional Non-Destructive Testing (NDT) techniques, which are based on the heuristic experiencebased echo-dynamic pattern identification methods, bring about costly, lengthy and error-prone analysis and thus lead to inconsistencies in results.

To address such a need, industry demands new innovative NDT techniques for shaft-typed steel pieces, and furthermore requires novel algorithms for the analysis of large volumes of ultrasonic shaft test data. More specifically, this paper aims to develop a more advanced and consistent Automatic Ultrasonic Signal Classification (AUSC) paradigm for testing shafts. Figure 1 highlights our comprehensive analysis and the research points corresponding to each stage of developing an AUSC system for testing shafts.

Section II is concerned with the extraction of informative features from ultrasonic signals, particularly focussing on two top approaches; namely, Fast Fourier Transform (FFT) and Discrete Wavelet Transform (DWT) both of which have been explored and compared each other by many researchers in their quest for better sets of features for AUSC. In Section III, the comparison analysis between two feature extraction schemes (FFT and DWT) is expanded through the intensive experiment of the classification performance by employing Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) as learning algorithms. We report results of our investigation into whether DWT can outperform FFT at extracting features for ultrasonic shaft test data. Section IV deals with an open issue: which generation and combination method to choose for constructing the most effective and reliable multi-model systems for our application domain. It suggests guidelines for the choice of ensemble structure for ultrasonic shaft signal classification through the experimental analysis. Conclusions are followed in Section V.

II. Feature Extraction Schemes for Ultrasonic Signals

A. FFT coefficients using magnitude and phase

In their quest for better sets of features for AUSC, many researchers have commonly employed two different preprocessing techniques using coefficients of FFT and coefficients of DWT in order to extract feature sets, and compared the classification performance using each feature set. Most results of comparing FFT and

Manuscript received January 15, 2009; accepted February 15, 2009.

doi:10.4304/jcp.5.5.679-686

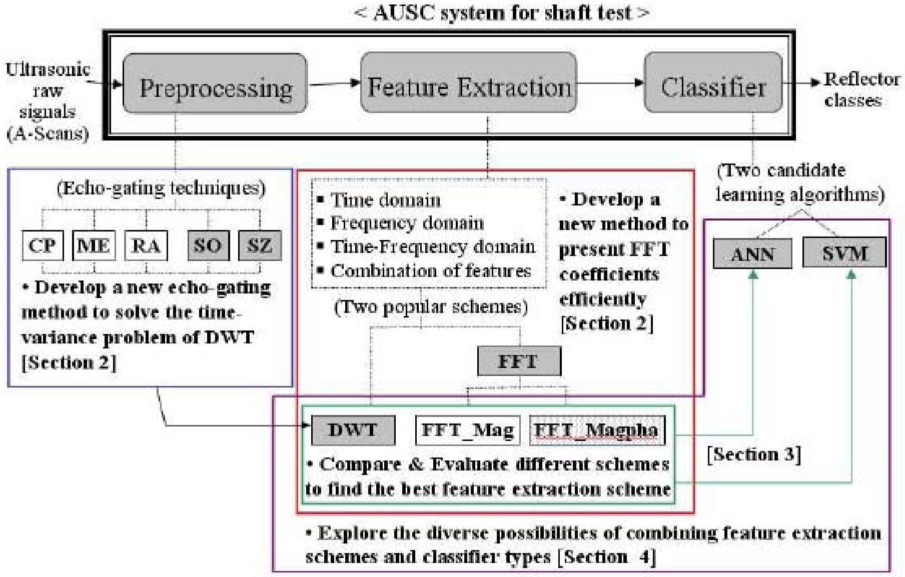

< AUSC system for shaft test >
Ultrasonic raw signals (A-Scans)
Preprocessing
Feature Extraction
Classifier
Reflector classes
(Echo-gating techniques)
CP
ME
RA
SO
SZ
• Develop a new echo-gating method to solve the time-variance problem of DWT [Section 2]
• Time domain
• Frequency domain
• Time-Frequency domain
• Combination of features
(Two popular schemes)
• Develop a new method to present FFT coefficients efficiently [Section 2]
FFT
(Two candidate learning algorithms)
ANN
SVM
DWT
FFT_Mag
FFT_Magpha
• Compare & Evaluate different schemes to find the best feature extraction scheme
[Section 3]
• Explore the diverse possibilities of combining feature extraction schemes and classifier types [Section 4]

Figure 1. The categorization of our research aims corresponding to each stage of developing an AUSC system.

DWT showed a superiority of DWT over FFT in discriminating the type of flaw (or its non-existence) in the context of comparatively shorter and simpler signals [4–6]. However, those previous claims are subject to debate because most previous comparison studies used only the magnitude component of the transformed signals using FFT and their phase components were naturally excluded through the process of using FFT coefficients as feature vectors for classifiers (we named this type of FFT coefficients "FFT_Mag"). Therefore, in order to make a fairer comparison, we formed a new set of FFT feature vectors which effectively represent both magnitude and phase information of FFT sequences (we named this type of FFT coefficients "FFT_Magpha").

A flowchart showing the process of constructing "FFT_Mag" and "FFT_Magpha" is presented in Figure 2. The final feature vector has the same number of elements because we apply down-sampling (by at least two) on the FFT sequence for the "FFT_Magpha" method. This adjustment of the length of the feature vectors is especially required for using them for the comparison experiment between two FFT schemes. Figure 3 presents the result of our experiment for comparing the classification performances of using "FFT_Mag" and "FFT_Magpha" respectively as feature vectors to ANN. The result implies that "FFT_Magpha" is a more efficient FFT based feature extraction approach than "FFT_Mag" thus confirms that the phase components of FFT sequences are important information carriers and must not be ignored. Therefore, in order to investigate whether DWT can outperform FFT as a feature extraction scheme in this application, it is more appropriate to compare the feature extraction schemes using DWT with the "FFT_Magpha" approach than with the "FFT_Mag".

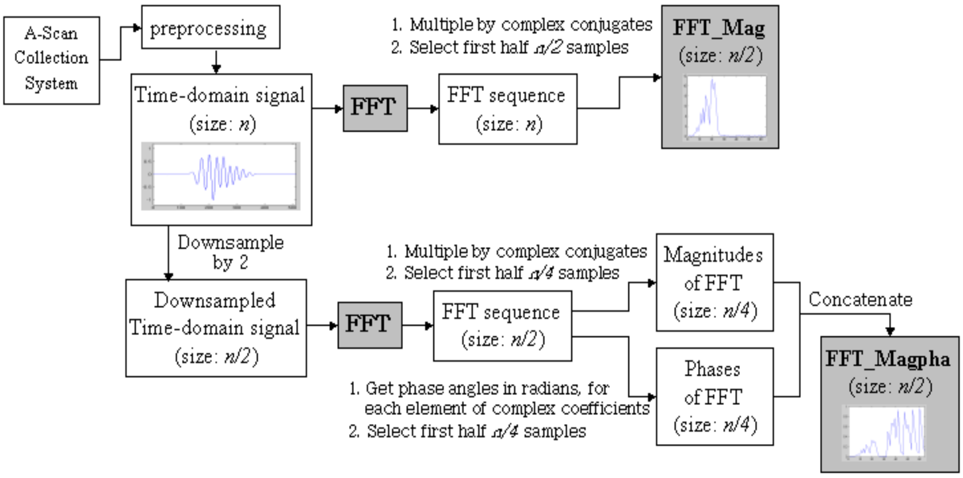

A-Scan Collection System
preprocessing
Time-domain signal (size: n)
FFT
FFT sequence (size: n)
1. Multiple by complex conjugates
2. Select first half n/2 samples
FFT_Mag (size: n/2)
Downsample by 2
Downsampled Time-domain signal (size: n/2)
FFT
FFT sequence (size: n/2)
1. Multiple by complex conjugates
2. Select first half n/4 samples
Magnitudes of FFT (size: n/4)
1. Get phase angles in radians, for each element of complex coefficients
2. Select first half n/4 samples
Phases of FFT (size: n/4)
Concatenate
FFT_Magpha (size: n/2)

Figure 2. The procedure of constructing FFT_Mag and FFT_Magpha

B. *DWT coefficients preprocessed using a new echo-gating technique*

Many of the recent work on ultrasonic flaw classification employed the DWT as part of its feature extraction scheme [4–8] and demonstrated the potential of DWT as a useful feature for AUSCs. DWT, however, exhibits a time-variance problem that has resulted in reservations about its wide acceptance [9-10]. To overcome this problem, we developed new techniques (SO[1] and SZ[2]) to derive a preprocessing method for time-domain A-Scan signals.[3] This techniques offer consistent extraction of a segment of the signal from long signals that occur in the NDT of shafts. It has been investigated if the newly developed method for gating a signal section and singling out an echo plays an effective role in overcoming the timevariance problems of DWT [11]. This investigation also included the comparison of the performance of the newlydeveloped technique (SO and SZ) with other alternatives (RA, CP and ME)[4], and established experimentally that DWT coefficients can be used as a feature extraction scheme more reliably by using our new preprocessing technique.

III. FFT *VS* DWT USING ANN OR SVM CLASSIFIER

Once the feature extraction process has been completed, a suitable decision making algorithm for classification is applied to determine and classify the flaw type information. Among the various learning mechanisms for ultrasonic signal classification, ANNs have gained more popularity due to their ability to generate complex decision boundaries in the multidimensional feature space [12]. This is attractive, especially in ultrasonic flaw classification, because the relationship between ultrasonic signal characteristics and their defect class is not straightforward [1, 13].

In order to evaluate which feature extraction scheme is best for an AUSC system for shaft testing using ANNs as its classification algorithm, we investigated into whether DWT can outperform FFT at extracting features in ultrasonic signals from shafts. Other previous reports [4, 14] have compared the DWT based features with the FFT with limited feature components. Typically, those previous reports considered only short signals and they paid little attention to the phase components of FFT sequences. As explained in Section II, we approached to represent both the magnitude and phase of the FFT sequences in a feature vector and compared these newly proposed FFT based feature sets (FFT Magpha) not only with the conventional FFT feature sets (FFT Mag) but also with DWT based feature sets.

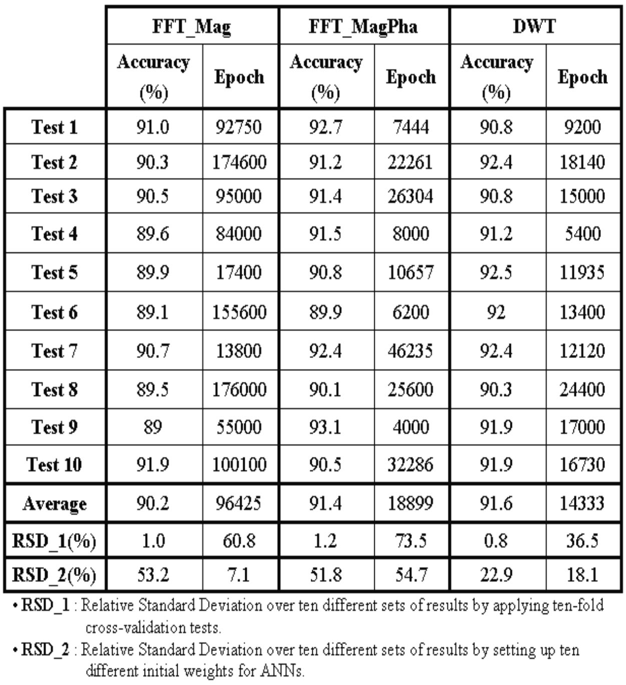
More precisely, three feature extraction schemes (FFT Mag, FFT Magpha and DWT) were used to represent the signal data feature and all parameters of the ANN classifiers were kept constant to focus on the feature extractions scheme. Thus, the data features have the same number of components and are presented to ANNs with the same architecture and parameters. The effectiveness of all feature extraction schemes were compared by their classification accuracy for ultrasonic shaft signal echoes. To test the performance of each feature extraction scheme, we recorded three result sets. Each result set corresponds to applying FFT Mag, FFT Magpha and DWT, and it is derived from each of the ten times of 10-fold cross validation tests (a total of one hundred tests for each network). The derived information consists of the following indicators.

1) The percentage of correct classification over the validation set.
2) The number of epochs required to train to the given error rate or to the lowest validation error rate before overfitting happens.

Table I lists the result values for each cross validation test set. The presented values for each test (from the 1st row to the 10th row of Table I) are the averages calculated for ten runs of test by setting up different initial weights on ANNs. An overall average is also calculated and put on the 11th row of Table I. We also calculated the corresponding standard deviation divided by this mean value (relative standard deviation) both over cross validation test sets and over different weighting runs for each test sets. They are presented on this table as RSD_1 and RSD_2.

The visual comparison between the classification

TABLE I
ACCURACY AND NUMBER OF REQUIRED EPOCHS FOR THREE DIFFERENT FEATURE EXTRACTION SCHEMES: FFT_MAG *VS* FFT_MAGPHA *VS* DWT



	FFT_Mag		FFT_MagPha		DWT	
	Accuracy (%)	Epoch	Accuracy (%)	Epoch	Accuracy (%)	Epoch
Test 1	91.0	92750	92.7	7444	90.8	9200
Test 2	90.3	174600	91.2	22261	92.4	18140
Test 3	90.5	95000	91.4	26304	90.8	15000
Test 4	89.6	84000	91.5	8000	91.2	5400
Test 5	89.9	17400	90.8	10657	92.5	11935
Test 6	89.1	155600	89.9	6200	92	13400
Test 7	90.7	13800	92.4	46235	92.4	12120
Test 8	89.5	176000	90.1	25600	90.3	24400
Test 9	89	55000	93.1	4000	91.9	17000
Test 10	91.9	100100	90.5	32286	91.9	16730
Average	90.2	96425	91.4	18899	91.6	14333
RSD_1(%)	1.0	60.8	1.2	73.5	0.8	36.5
RSD_2(%)	53.2	7.1	51.8	54.7	22.9	18.1

- **RSD_1** : Relative Standard Deviation over ten different sets of results by applying ten-fold cross-validation tests.
- **RSD_2** : Relative Standard Deviation over ten different sets of results by setting up ten different initial weights for ANNs.

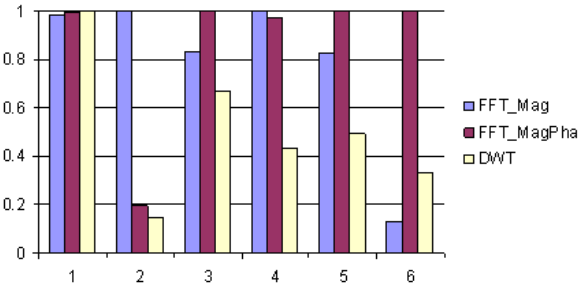
[1] Systematical echo capturing method with preservation of original neighbouring grass.

[2] Systematical echo capturing method with zero-padding.

[3] The details about these two techniques are presented in the author's previous work [11]

[4] RA: Random Positioning; CP: Central-Peak positioning; ME: Main Energy capturing.

performance of using three different feature extraction schemes is offered by Figure 3, which shows histograms presenting relative values of the results shown in Table I.



1. The percentage of correct classification over test data
2. Number of epochs required.
3. Relative standard deviation of the classification results for the 10 test sets used in the cross validation test.
4. Relative standard deviation of the classification results for the 10 sets of test by setting up different initial weights on ANNs.
5. Relative standard deviation of the required epochs for the 10 test sets used in the cross validation test.
6. Relative standard deviation of the required epochs for the 10 sets of test by setting up different initial weights on ANNs.

Figure 3. Relative comparison of performance of ANNs using three different feature extraction schemes: FFT_Mag *vs* FFT_Magpha *vs* DWT

In order to judge the statistical significance of the test results, we conducted ANOVA [15], [16] tests for the results at the p=0.05 level and the results are summarized in Table II. The statistical significance of the results is determined by comparing the F value produced through F-test with its corresponding F-test critical value. That is, if F value is bigger than the critical value, the evidence of statistical significance is produced and then the test results can be supported. The results from our experiments with three different feature sets suggest the following pair wise comparison.

TABLE II
A SUMMARIZED RESULT OF THE ANOVA TEST CONDUCTED FOR THE RESULTS PRESENTED ON TABLE I

Groups		**F-test**	
Feature Extraction Method	**Result**	**F value**	**F-test Critical Value**
FFT_Mag, FFT_MagPha, DWT	epoch	17.47	3.354
	accuracy	6.94	

First, does phase information count for something on long signals? That is, using only the magnitude of the FFT sequences as a feature set rather than using the magnitude and phase together as FFT based feature sets (but reduced sampling rate). Because we separated the magnitudes component and phase component of complex FFT sequences and rearranged them by concatenation, we can compare this feature extraction technique (FFT_Magpha) with the more generally used FFT based feature extraction technique (FFT_Mag). Secondly, does DWT buy more information than phase information? That is, a comparison between feature extraction schemes using FFT and DWT. From the analysis of the results the following conclusions can be drawn.

- Our experiments show that the FFT_Magpha provided the better result in classification performance. This experimental result is also supported by its ANOVA test result in Table II. Moreover, clearly a smaller number of epochs is required for convergence (for equivalent error tolerance) in the validation set when using the FFT Magpha scheme.
- This result implies that FFT Magpha is a more efficient FFT based feature extraction approach than FFT_Mag. If the “phase” components of FFT sequences, which can usually be ignored in the process of extracting the frequency components, are represented in the FFT based feature vectors, the precision as well as their training and execution time are highly enhanced.
- Therefore, in order to investigate whether DWT can outperform FFT as a feature extraction scheme in this application, it is more appropriate to compare the feature extraction schemes using DWT with the FFT_Magpha approach than with the FFT Mag. We can also say that FFT Magpha provides the benchmark for the FFT-related feature extraction scheme.
- With respect to classification accuracy, the DWT_based feature extraction method also provides results as good as the FFT Magpha method.
- The DWT also showed more reliability by producing comparatively stable results in different runs of cross validation tests. This implies that the DWT has potential as feature extraction scheme for training ANNs with arbitrary training data and using the networks for in-field ultrasonic shaft signal classification.

Though the demonstrated superiority of DWT was initially tested on ANNs as the classifier, this raises the issue about the synergy created by the DWT as a feature selector and the ANN as the classifier; namely, is this a feature extraction that is too much fit for ANN and not useful for other classifiers? That is, can the superiority of DWT be validated by the comparison of its classification performance with FFT’s only through ANN classifiers? Though ANNs have been popularly employed as classifiers for AUSC, we also need to consider the many difficulties inherent in the ANN learning paradigm (such as generalization control, overfitting and parameter tuning) thus should be careful in claiming DWT’s predominance.

In order to confirm the potential of DWT as an efficient feature extraction scheme for ultrasonic shaft test signals, we made a new comparative experiment involving SVM approach instead of ANN models. SVM has gained a strong reputation for its generalization control capability, thus avoiding overfitting, and more confidence can be placed in comparison results using SVM modelling than in those using ANN modelling, especially when there are only a limited number of training examples. We also analysed the classification results from both schemes (FFT and DWT) to investigate whether different feature extraction schemes affect the classification performance in different classes. That is, we recorded the classification accuracy for each of 10 test runs as well as the average accuracy, and also analysed how many instances of each class (CR and MC) were classified correctly. We summarize the classification results (as accuracy percentages) in Table III. Figure 4 offers a visual comparison between the classification performance of the two different feature extraction schemes. It displays the corresponding histograms of relative values for the results shown in Table III. In order to judge the statistical significance of the test results, we also conducted ANOVA tests for the results at the p=0.05 level and the test results are summarized in Table IV.

TABLE III
ACCURACY CLASSIFIYING EACH CLASS (CR AND MC) FOR TWO DIFFERENT FEATURE EXTRACTION SCHEMES FFT *VS* DWT

	FFT			DWT		
	CR	MC	Overall	CR	MC	Overall
Test 1	57.7	100	79.4	92.3	78.4	84.1
Test 2	83.3	100	86.7	90.2	75	88
Test 3	60.7	94.3	79.4	92.9	85.7	84.1
Test 4	57.9	90	61.9	100	82	87.5
Test 5	72.5	91	80	81	90	84.4
Test 6	62.5	100	82.5	91.7	84.6	87.3
Test 7	73.1	91.9	84.0	100	86.5	92.1
Test 8	63.6	93.5	81	100	75	81.3
Test 9	76.2	72.8	75	90	100	93.8
Test 10	64.3	91.7	79.7	92.9	77.8	84.4
Average	67.2	92.5	79	93.1	83.5	86.7

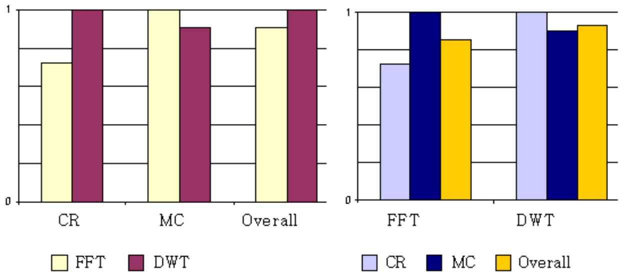


Figure 4. Relative comparison performance using two different feature extraction schemes: FFT *vs* DWT

TABLE IV
A SUMMARIZED RESULT OF THE ANOVA TEST CONDUCTED FOR THE RESULTS PRESENTED ON TABLE III

Groups		F-test	
Feature Extraction Method	Result	F value	F-test Critical Value
FFT_Mag, FFT_MagPha, DWT	epoch	17.47	3.354
	accuracy	6.94	

Through the analysis of the experimental results, we gained several noteworthy points as follows:

- The SVM classification performance of DWT is superior to that of FFT in every different test run. This matches the result of the previous comparison applying ANN models. It implies that DWT is a more reliable feature extraction scheme; that is, it can be employed for constructing a better classifier for in-field ultrasonic shaft signal tests. The result also dissipate any doubt that the DWT feature extraction methodology is too far suited for ANN.
- We can observe some differences for specific classes of echoes when reflecting upon the classification result from using each feature extraction scheme. DWT shows better performance in classifying CR than MO, while FFT shows a preference for the other way around. That is, using DWT one rarely gets a CR incorrectly classified (either a CR as something else or something else as a CR). Symmetrically, using FFT one rarely gets an MO incorrect.
- This result can provide useful prior knowledge when constructing a hybrid AUSC system for testing shafts using a combined feature extraction scheme. In this system, the FFT, in spite of lower accuracy for overall classification, could complement the decisions based on DWT features.

IV. ENSEMBLES FOR ULTRASONIC SIGNAL CLASSIFIERS

While we have focussed on finding the best single classifier model by determining the best selected feature extraction scheme (FFT or DWT), we also try to learn multiple models of the shaft test data and combine their outputs for making a final decision for classification. The reason for combining multiple models (*ensemble of classifiers*) constructed by a single feature extraction scheme (FFT or DWT) and a single learning algorithm (ANN or SVM) is that FFT might reflect physical properties that are different from those that DWT shows. We suspect that including the FFT as another informant of the decision process, even if the accuracy using DWT has shown to be superior, should improve accuracy.

There are various approaches (homogeneous or heterogeneous) for generating multiple classifiers and for combining the outputs of multiple models [17-18]. The issue raised by this diversity of methods for generating and combining multiple models is to pursue the best generation-combination method for constructing the most effective and reliable multi-model AUSC system for our application domain. Therefore, in order to construct an effective multi-classifier system, we need to decide a scheme for creating models and also a combination method for decision making.

To determine guidelines for the construction of an integrated multi-classifier model using both feature schemes (FFT and DWT) effectively. we explored the diverse possibilities of heterogeneous and homogeneous ensembles, combining techniques, feature extraction methods and classifier types. The whole experimental setting consists of the following steps

1) Map shaft inspection data into feature domains using two feature extraction schemes (FFT and DWT).
2) Train them through SVM models and ANN models using five-fold cross-validation learning and recorded their performance as single models.
3) Combine single models across two dimensions: a) combining the decisions of the FFT model and the DWT model trained by a single learning paradigm and b) combining the decisions of the SVM model and the ANN model with the same feature scheme. We applied three different traditional combining methods (DS, BC and LC) for each combination.
4) Compare the classification accuracy results from the three combined models with the results of using a single model.

Note that the first and second steps constitute procedure for generating multiple models. The third step is the procedure for combining those multiple models. Figure 5 graphically summarizes those two procedures (model generation and model combination) of our experiment and Figure 6 presents the whole procedure in our experiment in summary.

In order to investigate if the combined model performs more accurately in classifying input data than a single model does, we compared the decision error rate of the combined model with the decision error rate of each individual participant model. Table V shows the comparison of the decision error rate between combined models and single classifier models by calculating the difference of both error rates. $c(A,B)$ indicates a combined classifier of two individual classifiers model A and model B. The figures in rows (1) to (3) indicate the difference of error rates between the combined model and the first single model participant while those in (4) to (6) are the difference of error rates between the combined model and the second participant. Among these six rows, (1) and (4) are the results for classification performance for overall data. On the other hand, (2) and (5) are result figures for classifying CR while (3) and (6) are for classifying MC.

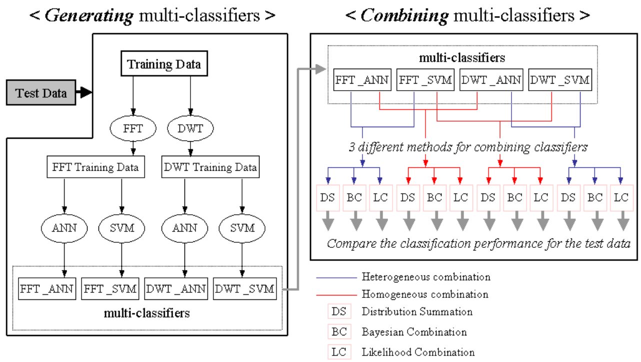


Figure 5. Diagram of our experiment presenting two main procedures: model-generation and model-combination

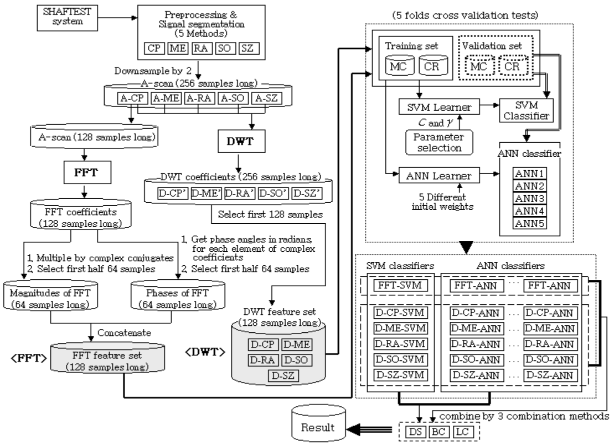

SHAFTEST system
Preprocessing & Signal segmentation (5 Methods)
CP
ME
RA
SO
SZ
Downsample by 2
A-scan (256 samples long)
A-CP
A-ME
A-RA
A-SO
A-SZ
A-scan (128 samples long)
DWT
FFT
DWT coefficients (256 samples long)
D-CP'
D-ME'
D-RA'
D-SO'
D-SZ'
FFT coefficients (128 samples long)
Select first 128 samples
1. Get phase angles in radians for each element of complex coefficients
2. Select first half 64 samples
1. Multiple by complex conjugates
2. Select first half 64 samples
Magnitudes of FFT (64 samples long)
Phases of FFT (64 samples long)
DWT feature set (128 samples long)
Concatenate
<FFT>
FFT feature set (128 samples long)
<DWT>
D-CP
D-ME
D-RA
D-SO
D-SZ
(5 folds cross validation tests)
Training set
MC
CR
Validation set
MC
CR
SVM Learner
SVM Classifier
C and γ
Parameter selection
ANN classifier
ANN Learner
5 Different initial weights
ANN1
ANN2
ANN3
ANN4
ANN5
SVM classifiers
ANN classifiers
FFT-SVM
FFT-ANN
FFT-ANN
D-CP-SVM
D-ME-SVM
D-RA-SVM
D-SO-SVM
D-SZ-SVM
D-CP-ANN
D-ME-ANN
D-RA-ANN
D-SO-ANN
D-SZ-ANN
D-CP-ANN
D-ME-ANN
D-RA-ANN
D-SO-ANN
D-SZ-ANN
Result
combine by 3 combination methods
DS
BC
LC

Figure 6. Overall procedure of our experiment for building an ensemble of classifiers

Figure 7 graphically presents this comparison result, and displays the amount of improvement in the classification performance in a form of bar charts. The Bars where the combination is an improvement point downwards while if a single classifier remains better, the bar points upwards. We also computed a value ϕe which indicates the “fraction of correlated errors” [19] and is also listed in Table V and Figure 7. The value of e is generally used to measure the degree to which the errors made by models of the ensemble are correlated.

The following points are remarkable from this experimental result.

- Combined models show better performance than single models in the classification accuracy for the whole test set across schema for generating or combining multi-classifiers.
- Combining two classifiers trained by different feature sets becomes more advantageous when we use SVM as a learning algorithm than when we use ANN.
- Though the overall accuracy of combined models is higher than the accuracy of single models across most types of combinations, their performance in classifying each class data (MO and CR) is diverse. Especially, most FFT and DWT ensembles trained by ANN perform worse than single models in classifying CR data; whilst corresponding combined models trained by SVM perform reliably on both classes except for one combination (the FFT and DWT-CP ensemble).
- Amongst the five types of DWT data combined with FFT data, DWT-SZ shows most reliability in classifying both classes regardless of the learning paradigm. This implies that different echo gating preprocessing for extracting DWT features plays a role in structuring the DWT feature sets. We

suspect there are some implicit differences in DWT.

- The performance of the heterogeneously combined classifiers is different depending on which feature sets were used to train them.
- The value of ϕe is related to the amount of error reduction made by combining multi-classifiers. As shown in Figure 7, the value of ϕe seems to be very relevant to the overall error reduction rate but not to have much relevance to the error reduction for each class data.
- The most suitable combination structure may depend on the interest of some particular class. For example, if accuracy for the CR class is the issue, then the SVM with DWT (single classifier) is not surpassed by the combination, although the combination does better over all the classes.

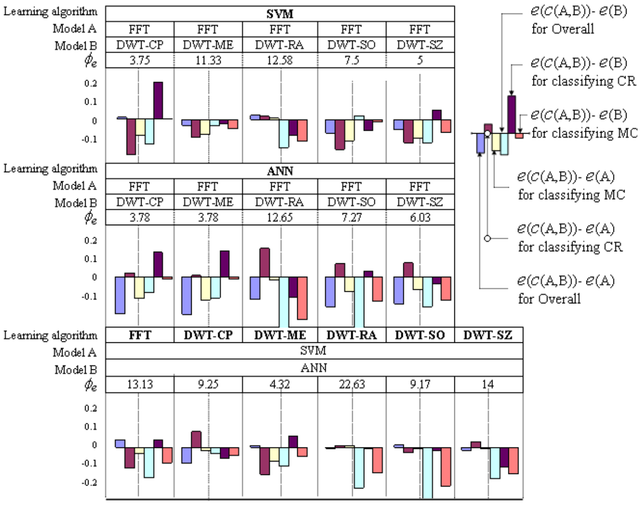


Figure 7. Bar charts for the comparison of the performance of combined models. *c(A,B)* indicates a combined classifier of two individual classifiers, model A and model B. *e(A)* indicates the classification error rate of the classifier model A.

V. Conclusions

The industrial application discussed in this study is the classification of ultrasonic echoes in an A-scan. The application is particularly challenging as A-scans are taken from the end of long large complex metal shafts. The problem is then, to discriminate efficiently the different types of reflectors amongst the large volumes of digitalized ultrasonic shaft defect information. This paper reports on our experimental investigation on efficient feature extraction schemes and classifiers for shaft testing system. The following is a list of contributions made in this study.

- Introduction of a new FFT-based feature extraction scheme(FFT_Magpha) by which both magnitude and phase components of FFT sequences are effectively represented. Through this newly developed feature extraction scheme, we can incorporate the "phase" component of FFT sequences (which can, in other schemes, usually be ignored in the process of extracting the frequency components) to the construction of the FFT-based feature vectors. The experimental comparison between this new scheme and the traditional scheme (FFT_Mag) by using them in a typical AUSC system, shows that the new scheme (FFT_Magpha) highly enhances the precision as well as improving their training and execution time.
- Analysis of DWT as a more beneficial feature extraction scheme than FFT in an AUSC system for testing shafts. This is analysed through the experimental comparison of the classification performance using DWT, not only with the classification performance using the traditionally-used-FFT with limited feature components, but also with the newly-proposed-FFT scheme (FFT_Magpha). This extended comparison between DWT and the state-of-the-art FFT provides a more reliable and trustworthy analysis about DWT as a feature extraction scheme for our application.
- Finding the potential of DWT as a more reliable feature extraction scheme, through the more stable classification results in different runs of cross validation tests compared to the results produced in the tests using FFT-based feature extraction scheme. This potential is especially beneficial for the practical NDT for shafts, as we can train a classifier with arbitrary training data and then use the classifier for in-field ultrasonic shaft signal tests.
- Demonstration of the superiority of using DWT as the feature extraction scheme in the ultrasonic shaft signal classification involving not only ANN but also SVM. These results dissipate any doubt that the DWT feature extraction methodology is too far suited for ANN which has been popularly employed previously in many similar experimental scenarios.
- Discovery of predisposition to distinguish a certain facility when specific classes of echoes are concerned with different combinations of feature extraction (FFT or DWT) and classifier (ANN or SVM), though DWT is superior to FFT and SVM is superior to ANN in terms of the overall classification accuracy. This finding leads into a hybrid classifier that will improve overall performance by giving more weight to the more trustworthy sub-classifier.
- Exploration of the diverse possibilities of heterogeneous and homogeneous ensembles by combining classifiers along the dimension of feature extraction mechanism, along the dimension of combination methods and along the dimension of type of classifier

REFERENCES

[1] G. Katragadda, S. Nair, and G. P. Singh, "Neuro-Fuzzy Systems in Ultrasonic Weld Evaluation," *Review of Progress in Quantitative Nondestructive Evaluation*, vol. 16, pp. 765–772, 1997.

[2] S. J. Song, H. J. Kim, and H. Lee, "A systematic approach to ultrasonic pattern recognition for real-time intelligent flaw classification in weldments," *Review of Progress in Quantitative Nondestructive Evaluation*, vol. 18, pp. 865–872, 1999.

[3] G. Cotterill and J. Perceval, "A New Approach to Ultrasonic Testing of Shafts," in *Proceedings of the 10th Asia-Pacific Conference on Non-Destructive Testing (APCNDT)*, [Online], 2001.

[4] R. Polikar, L. Udpa, S. S. Udpa, and T. Taylor, "Frequency Invariant Classification of Ultrasonic Weld Inspection Signals," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 45, no. 3, pp. 614–625, May 1998.

[5] D. Redouane, K. Mohamed, and B. Amar, "The Investigation of Artificial Neural Network Pattern Recognition of Acoustic Emission Signals for Pressure Vessel," in *Proceedings of the 15th World Conference on Non-Destructive Testing*, [Online], 2000.

[6] J. Spanner, L. Udpa, R. Polikar, and P. Ramuhalli, "Neural networks for ultrasonic detection of intergranular stress corrosion cracking," *The e-Journal of Nondestructive Testing And Ultrasonics*, vol. 5, no. 7, [Online], July 2000.

[7] M. S. Obaidat, M. A. Suhail, and B. Sadoun, "An intelligent simulation methdology to characterize defects in materials," *Information Sciences*, vol. 137, pp. 33–41, 2001.

[8] G. Simone, F. C. Morabito, R. Polikar, P. Ramuhalli, L. Udpa, and S. Udpa, "Feature extraction techniques for ultrasonic signal classification," in *Proceedings of the 10th Int. Symposium on Applied Electromagnetics and Mechanics (ISEM 2001)*, 2001.

[9] Y. Mallet, D. Coomans, J. Kautsky, and O. D. Vel, "Classification Using Adaptive Wavelets for Feature Extraction," *IEEE Transactions of Pattern Analysis and Machine Intelligence*, vol. 19, no. 10, pp. 1058–1066, October 1997.

[10] S. Pittner and S. V. Kamarthi, "Feature Extraction From Wavelet Coefficients for Pattern Recognition Tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 1, pp. 83–88, January 1999.

[11] K. Lee and V. Estivill-Castro, "Feature extraction and gating techniques for ultrasonic shaft signal classification," *Applied Soft Computing*, vol. 7, no. 1, pp. 156–165, 2007.

[12] F. W. Margrave, K. Rigas, D. A. Bradley, and P. Barrocliffe, "The use of neural networks in ultrasonic flaw detection," *Measurement*, vol. 25, pp. 143–154, 1999

[13] J. B. Santos and F. Perdig~ao, "Automatic defects classification – a contribution," *NDT & E International*, vol. 34, pp. 313–318, 2001.

[14] S. V. Kamarthi and S. Pittner, "Fourier and Wavelet Transform for Flank Wear Estimation - A Comparison," *Mechanical Systems and Signal Processing*, vol. 11, no. 6, pp. 791–809, 1997.

[15] G. Clarke and D. Cooke, *A Basic Course in Statistics*. Arnold, 1998.

[16] P. Cohen, Empirical Methods for Artificial Intelligence. MIT Press, 1995.

[17] T. G. Dietterich, "Machine-learning research: Four current directions," *AI Magazine*, vol. 18, no. 4, pp. 97–136, 1997.

[18] G. Valentini and F. Masulli, "Ensembles of Learning Machines," in *Neural Nets WIRN vietri-2002*, ser. Lecture Notes in Computer Science 2486, R. Tagliaferri and M. Marinaro, Eds. Springer-Verlag, 2002, pp. 3–19.

[19] K. M. Ali and M. Pazzani, "Error reduction through Learning Multiple Descriptions," *Machine Learning*, vol. 24, no. 3, pp. 173–202, 1996.

.

Kyungmi Lee received her Ph.D. in School of Computing and Information Technology from Griffith University, Australia in 2007, her Graduate Certificate from School of Electrical Engineering and Computer Science from the University of Newcastle, Australia in 2001, and her BSc in Computer Science from the Pusan National University, Korea in 1993. Dr. Lee joined Charles Stuart University after her Ph.D. as lecturer. Recently, she joined School of Business in James Cook University as a lecturer. She has been actively involved in collaborative research projects in the areas of machine learning, non-destructive signal processing, geo-informatics and business intelligence. Dr. Lee has been involved in some academic activities including reviewing articles for journals and conferences, serving as a member of program committee for machine learning related conferences, and supervising students.

Design of Grid Resource Management System Based on Information Service

ZHANG Qian
College of Computer and Communication Engineering, China University of Petroleum
Dongying, Shandong, China 257061
Email:zhangqianupc@163.com

LI Zhen
College of Computer and Communication Engineering, China University of Petroleum
Dongying, Shandong, China 257061
Email:lizhen_upc@163.com

Abstract—**In this paper, a design-oriented information service grid model for resource management and to achieve the various components of the function. The author presents an improved "divided" algorithm based on the traditional, classical Min-Min scheduling algorithm, and then simulates the algorithm by GridSim Grid simulator. Comparing with the traditional Min-Min algorithm, this improved one overcomes the disadvantage of load balancing preferably. A kind of fault tolerance scheduling tactics with real-time characteristic was introduced and implemented in this paper.**

***Index Terms*—resource management system, grid, job scheduling, Divided-Min-Min, fault tolerance scheduling.**

I. INTRODUCTION

Grid resource management is an important component of grid system, its main function is to identify resource requirement, match and allocate resource, schedule and monitor resource, so as to use resource efficiently as far as possible. Grid resource management focus on controlling grid resource how to provide capacity and available service to other petitioners, its greatest feature is the virtualization and coordinated use of distributed and heterogeneous resource. It is precisely because the features exist in grid such as distributed, heterogeneous and dynamic, making resource management on grid more complex than that of clusters or distributed computing environment, so it is very necessary to establish a resource management system model adapts to grid, research its features and functions for achieving specific grid resource management system. At present the majority of the research is only limited to the theoretical or prototype system, the manager has been achieved mainly include GRAM (Grid Resource Allocation and Management), Condor and so on. But GRAM provides only a basic management framework, there is no specific algorithm, and Condor is actually a local grid scheduler, can not across VO (Virtual Organization) Dispatch. This paper designs a model of grid resource management system, and proposes a new higher performance resource scheduling algorithm on the basis of the original research.

II. MODEL OF GRID RESOURCE MANAGEMENT SYSTEM

Grid resource management is the core component of grid, the structure of grid resource management system designed in this paper is shown as Figure 1 in the next page. The system is built on GT4, with a hierarchical structure, and it provides users with such functions as resource discovery, job submission, job management and monitoring. The scheduling model consists of job collector, scheduler, manager, information collector and database.

A. *Job Collector*

Job collector is a user-oriented interface in the entire global scheduler. Whether the job is submitted through portal, or application program, the user needs to provide job name, location, necessary parameters for execution, name and path of the output file and other information. Job collector collects the information and stores them in database serviced for future job scheduling.

B. *Information Service*

The main function of Grid Information Service is to integrate all kinds of information in grid system, static and dynamic, shield heterogeneous nature, and provide a unified information access interface for users. In distributed, heterogeneous and scalable grid environment, information service not only needs to integrate the static system information (such as operating system, CPU, disk and memory size and other information), but also provide a number of dynamic update information such as

doi:10.4304/jcp.5.5.687-694

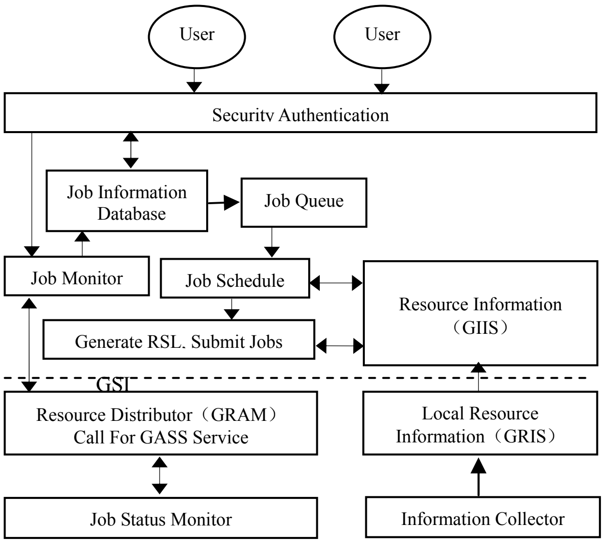

User
User
Securitv Authentication
Job Information Database
Job Queue
Job Monitor
Job Schedule
Resource Information （GIIS）
Generate RSL. Submit Jobs
GSI
Resource Distributor（GRAM） Call For GASS Service
Local Resource Information（GRIS）
Job Status Monitor
Information Collector

Figure1. Architecture of resource management system

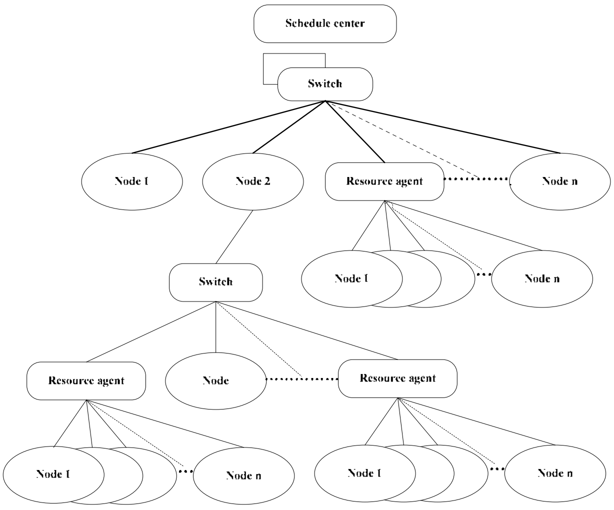


Figure2. Tree-level topology

the load and so on.. Information service mainly includes two parts, resource acquisition and resource evaluation.

1. Tree-level Topology

Network topology is the definition of technology to construct the physical grid environment [16], it is the objective need and precondition to establish resource model. Based on the classification scheduling model, this paper establishes a tree-level network topology, as shown in Figure 2. Resource agents and computing resources to be released constitute a multi-layered tree structure, including a primary resource agent (control center), a number of resource sub-agents and computing resources registered in these agents. Primary resource agents are the root nodes of a tree, computing resources are the leaf nodes, and sub-agent can be sub-node of a root node or a sub-agent.

Resource agent is responsible for receiving the scheduling task from parent node, and the task will be distributed to all child nodes, when the task is completed, it will return the results to the parent node. In addition, the resource agent is also responsible for the management of the various sub-nodes to achieve the control, add, delete, and other functions, at the same time, monitoring the operation of scheduling and dispatching log records. Resource nodes execute tasks distributed by scheduling agent in their local area and return the results. It will report the node case, as well as the count and completion of jobs on this node to the resource agent at a certain time interval T (typically 5 seconds).

Because of the existence of a large number of resources and users in grid environment, taking the feasibility, efficiency and scalability issues into account, it is not realistic that using a single dispatching center to deal with all of the web user's requests in the entire web. Tree-level topology presented in this paper has given full consideration to the system scalability, so it can easily access resources sub-tree node or topology.

2. Resource Discovery

Access to information resources can be divided into two types, static obtaining and dynamic obtaining. Static obtaining means that the grid resource information which is fixed will be stored in a fixed document, and we can read the document to obtain the necessary information. Dynamic obtaining establishes a connection to a certain host with globus installed, uses information services provided by globus to obtain real-time, dynamic resource information. So, the module is divided into two parts. ① Dynamic access to host information: In the interaction module, the scheduler provides a top node or monitor node, after establishing a connection through the web service approach using the information service interface provided by globus, we can call information inquire function provided by MDS with xpath statement, to obtain the required host information, and put them into a XML file. ② Static access to host information: Read the host list from the documents provided, and corresponding to a host, there is a certain file to store host attributes, access to the file, and read from it for more information. Because the host information is based on the form of XML documents, the scheduler does not recognize them. So we should parse the XML documents to the format can be identified by Scheduler.

3. Resource Assessment

Resource assessment relates to the efficiency of scheduling, and there are two methods to be considered for resource assessment.

(1) Evaluate firstly and then schedule.

Before each scheduling, evaluate resource nodes under the jurisdiction, and then schedule in accordance with the scheduling strategy.

(2) Initialize firstly and then schedule, if it fails, update the resources and re-schedule.

At boot time, update the status of various resources firstly, and then submit and schedule. If it fails, re-evaluate the node, update its status, and at the same time re-schedule jobs according to the log file.

For the first approach, the time for generating scheduling T equals to the summation of the time for evaluating and the time for generate results in accordance with scheduling strategy. If we use single-

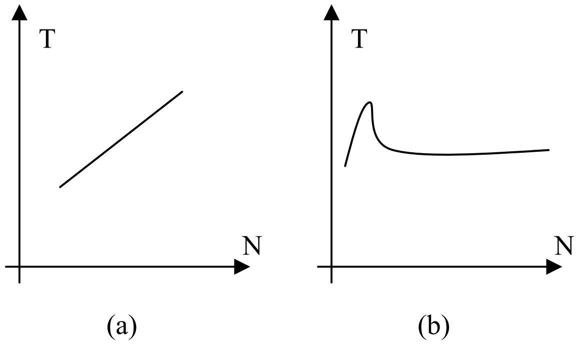


Figure3. The contrast simulation of two assessment ways

threaded to evaluate each node, the time will be the sum of each node, so this method has lower efficiency. Using a multi-threaded for evaluation, generate threads according to the number of nodes while scheduling, and implement all these threads at the same time for evaluating, so the assessment time will be shortened. Assessment time equals to the time spent by the thread received assessment information lastly. However, with the increase of node number, scheduling nodes need to generate many assessment threads, this will take considerable cost to run and maintain these threads, and lead to performance degradation for scheduling node, It is clear that such method is more suited to smaller nodes scheduling system.

The second approach is more suitable for large-scale system. This assessment draws on the solution to deal with deadlock problem in operating system. When the system starts up, evaluate all the registered resources nodes. When the system is initialized, each node in the resource pool will obtain the latest status information and wait for scheduling. In an assessment of each node, the use of multi-threading technology can shorten the system initialization time. In an assessment for each node, the use of multi-threading technology can shorten the system initialization time.

Figure 3 is the contrast simulation of these two resource assessment ways, and (a) represents the first one and (b) represents the second. Abscissa is the time to obtain scheduling results, as shown, the time to obtain scheduling results has different changes with the increase of the resource nodes number. Through the above analysis, this paper adopts the second assessment methods, and has achieved good results.

Resource information collector is targeted mainly on enquiry of the grid node and its information. Grid node information consists of static information and dynamic information. Static information does not change with time, such as hardware type, memory size, the type and version of the operating system, the information is obtained once at the time by sampling. Dynamic information is sampled with a fixed time interval, such as utilization of CPU and memory, and length of job queue.

As dynamic information plays a more important role for the grid resource scheduling, so it needs to ensure the real-time dynamic information. In this paper, MDS4, an information service component in GT4, achieves collection and dissemination of grid information.

C. *Job Scheduler*

The heterogeneous and dynamic features of grid system, and the different needs for resources, make resource scheduling as a very complex issue. There are two key points grid job scheduling facing, First of all, as a result of very large scale and due to the scalability reasons, it is hard to have an overall control center for scheduling, so it will face a range of issues such as load balancing during the process that the task is mapped to the available pool of resources. Secondly, the grid is dynamic, so scheduling should be adapted to changes in resources. Due to heterogeneous, wide-area, dynamic characteristics, a number of hardware errors, network errors and so can lead to failures of nodes or resources. So resource failures are quite common in grid. The above-mentioned errors must be detected and timely treated, and scheduler must have the corresponding fault-tolerant approach for errors associated with the application.

In response to these problems, the idea of scheduling model designed by this paper is mainly reflected in the classification scheduling, dynamic scheduling and fault-tolerant scheduling three aspects. Model uses a two-tier structure: global scheduler and local scheduler. Global scheduler is equivalent to the overall control center, and it is responsible for grid-wide resource discovery and scheduling. Local scheduler is equivalent to scheduling sub-agents, and it takes in charge of the domain scheduling. Jobs submitted by users are scheduled firstly in a control center, and secondly on the sub-agents. If there are more clusters or LANs following, there should be multi-level scheduling, such as the three-tier and the four-tier. Here we only consider two scheduling model. Model uses resources daemon processes to obtain and monitor information and load on all nodes in real-time, and update resource node information in the database at the same time to provide dynamic information for scheduling. When monitoring

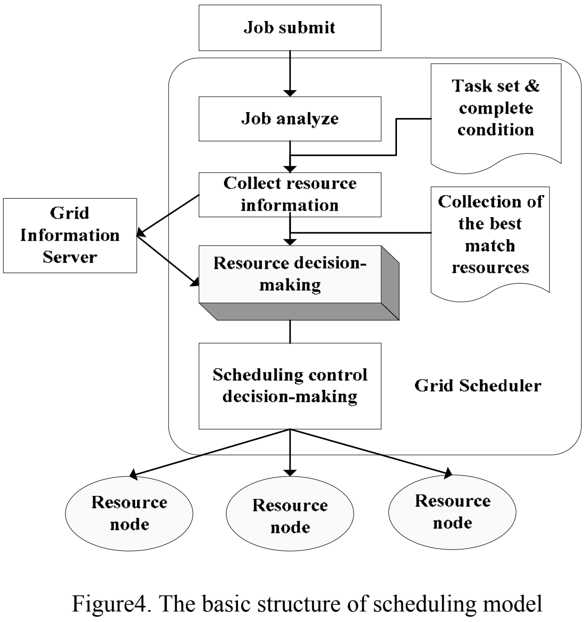

Job submit
Job analyze
Task set & complete condition
Collect resource information
Grid Information Server
Collection of the best match resources
Resource decision-making
Scheduling control decision-making
Grid Scheduler
Resource node
Resource node
Resource node

Figure4. The basic structure of scheduling model

the failure of resources, re-schedule the tasks through the fault-tolerant scheduling strategy. This scheduling system model draws on the simple and high-performance of centralized scheduling, as well as flexible organization and easy expansion of distributed scheduling, so it supports the scalability and fault tolerance, and can better adapt to the dynamic grid environment changes.

Based on the above design idea, we put forward a basic structure of scheduling model, as shown in Figure 4. It based on one scheduling agent for the tree topology, and can be applied to other scheduling agent with simple modifications.

D. Job Manager

Job manager consists two functions of job submit device and job monitor.

The roles of job submit are: (1) Schedule the job to resource node according to scheduler results. In order to achieve job distribution, scheduler needs to generate RSL file for job, and then submit it to a specific resource node. Meanwhile scheduler writes scheduling information in log file. (2) When the job completed, submit device also takes the responsibility of collecting results. Provide online result information regardless the job is success or failure.

As unsteadiness exists in local grid resource, users job may fail to execute, the role of job monitor is monitoring job status information, in order to deal with a fault-tolerant in time monitoring the failure job. Basic status of grid job including: Submitted (submitted), Pending (ready in scheduling queue), Running (running), Done (complete successfully), Failed (failure). Various conversions between the job states are: once job has been submitted, status shows as Pending, if execution starts up, the job will be deleted from Pending table and added to Running table, means the job is running. If job status is Done or Failed, delete the job from Running table, meanwhile update grid job information to the corresponding state.

E. Scheduling Log

Scheduling log is used to record scheduling result, ensure scheduling successfully. Each log record contains: scheduling mark (ID of schedule log); job mark (index of job for scheduling, NULL means new job); resource mark (index of resource distributed, generally for IP address); job description (describe job using RSL); job status (mainly include two item: submit and failure); submit time (the time when schedule occur);

As grid resource is dynamic and instability, fault or exit may occur on resource. When a resource node is found failed by resource monitoring, job scheduler will regenerate RSL request according to scheduling log to achieve scheduling result.

III. GRID JOB SCHEDULING ALGORITHM

Job scheduling system is the core part of grid resource management system, it receives job request, chooses appropriate resource to run the job. Job scheduling algorithm determines which scheduling policy to use for resource matching. In the absence of specialized scheduling algorithm for grid system, the vast majority of scheduling algorithms used in grid application have been using in distributed system or cluster system. The scheduling algorithms used commonly are Min-Min, Max-Min, fast greedy algorithm, Sufferage algorithm and so on.

A. Min-Min Scheduling Algorithm

Min-Min scheduling algorithm firstly calculate the minimum runtime of each job on all machines to select the job with minimum runtime and distribute it to appropriate host, then delete the job mapped recently from sets, and repeat this process until all the jobs have been mapped. As Min-Min algorithm maps small job (with minimum runtime) to fast machine, most small jobs will be assigned to machines which not only complete them first, but also have fast speed. However, large jobs with low priority will be assigned to less efficient machines, which cause load imbalance and lower utilization rate of available hosts in grid environment. For this, this paper proposes an improved method to schedule jobs with long local runtime firstly, then divides jobs into segment according to the job ETC value, so large job will be scheduled firstly, and the job in each segment will still be scheduled using Min-Min algorithm, this is Divided-Min-Min (hereinafter referred to Dmm) algorithm.

B. Divided-Min-Min Scheduling Algorithm

The essence of job scheduling problem is that, schedule m jobs T = (t1, t2,…, tm) to n hosts H = (h1, h2,…, hn) using a reasonable manner to gain the total runtime as small as possible (Makespan), under the grid environment composed by m jobs that needed to be scheduled and n available job execution unit (host or clusters). The expected runtime ETC (Expected Time to Compute) of m jobs on n hosts is a m*n matrix. ETC(i,j) shows expected runtime of the i-th job executing on the j-th machine, in the matrix, each row shows runtime of a certain job executing on each of n machines, each column shows runtime of m jobs executing on this host.

Dmm algorithm firstly sort jobs according to their ETC value, classify them as a sequence by average ETC value or minimum ETC value, or maximum ETC value. Then divide it into segments with the same size, and schedule large job segment firstly and then the small. For each segment, Min-Min algorithm also will be used for scheduling. Dmm algorithm described as follows:

(1) Calculate sort value key_i of each job. In heterogeneous grid environment, execution time of the same job in different machines is different, called grid job heterogeneity. Taking into account that, test three sub-strategies—the average, minimum, maximum expected execution time when determine sort value.

Sub-strategy 1 Dmm-avg: calculate average value of each row in the matrix.

$$\ldots \qquad key_i = \sum_j ETC(i,j)/m$$

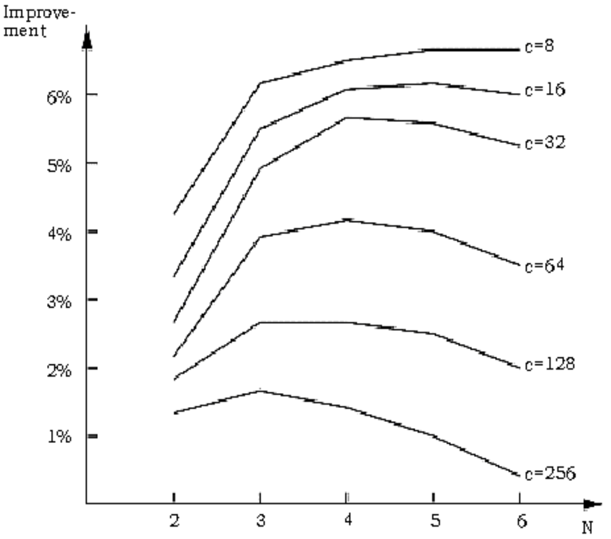


Figure 5. Effect of N value on algorithm improvement

Sub-strategy 2 Dmm-min: calculate minimum value of each row in the matrix.

$$\dots \qquad key_i = \min_j ETC(i, j)$$

Sub-strategy 3 Dmm-max: calculate maximum value of each row in the matrix.

$$\dots \qquad key_i = \max_j \; ETC(i, j)$$

(2) Arrange jobs by sort value in descending order to form a sequence.

(3) Divide job sequence into the average N segment. Focus on how to choose the best N value. On the one hand, the larger is N value the more balanced becomes the load, on the other hand, Min-Min algorithm will reduce efficiency with too large N value. The curve in Figure 5 shows improvement compared algorithm using Dmm-avg sub-strategy with Min-Min algorithm when select different N values. As shown in Figure 2, when the c=m/n value is small, that is the average number of jobs distributed to each machine is small, Dmm algorithm shows good performance.

Regardless of c value, the curve in Figure 2 always achieves the highest degree of algorithm improvement when N is 4, so this paper sets the N to 4 and usually divides job sequence into 4 segments.

(4) Schedule jobs of each segment in turn using Min-Min scheduling algorithm.

Different from Min-Min algorithm, Dmm algorithm sorts the job before scheduling, which means that the job with long execution time will be scheduled earlier. Then Min-Min algorithm will be used locally in each job segment. The key of this algorithm is how to determine the job order to ensure that priority scheduling of large jobs.

C. *Scheduling Algorithm Performance Test*

In order to test the performance of Dmm algorithm, this paper simulates 10 resource nodes and some jobs which consist of many sub-jobs. Test is composed by two groups of data and is compared with common Min-Min, Max-Min algorithm.

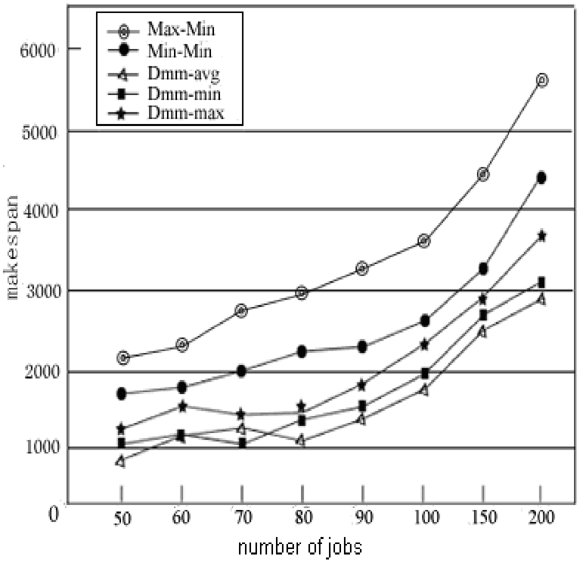


Figure6. Makespan comparison of five algorithms

1. Test makespan of algorithm

Figure 6 shows the simulation result using five different scheduling algorithms to 10 processors when numbers of grid job are different. Each point is the average value of five simulation results in this figure. As Figure, performance of three sub-strategies are better than the Max-Min and Dmm -min algorithm, in some cases Dmm-min performs better than Dmm-avg, but in most cases lower than Dmm-avg, and Dmm-max always performs lower than Dmm-avg. Therefore, this paper takes Dmm-avg as Dmm algorithm. Simulation data shows that, the performance of Dmm-avg algorithm increases 4.2 percent to 6.1 percent than Min-Min algorithm, and it increases 16.6 percent to 54.3 percent than Max-Min algorithm.

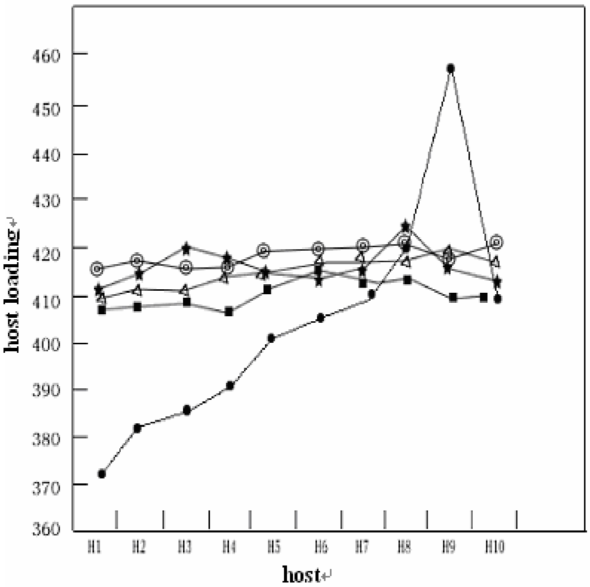


Figure7. Load balancing comparison of five algorithms

Experimental results show that the execution time of Dmm algorithm is shorter than Min-Min algorithm, because the Min-Min algorithm needs to search a job with the shortest execution time in the entire ETC matrix, and Sub-Dmm algorithm uses the subsection method, so it just needs to search in each segment. This method not only reduces Makespan, and shortens the execution time.

2. Test load balancing of algorithm

As shown in Figure 7, the load balancing of Dmm algorithm is better than the Min-Min algorithm and Max-Min algorithm. This is because Dmm algorithm schedules jobs with the shortest execution time in large job segment first, so most of the large jobs will be allocated to machines that not only complete them first, but also have faster execution speed, so the utilization of host running jobs is balanced relatively. Of three sub-strategies, the load curve of Dmm-avg is the smoothest, which means the algorithm is of the highest load balancing.

The future can be drawn through the above analysis, compared with the Min-Min scheduling algorithm, Dmm algorithm not only inherits good performance and simple character from it, but also has been noticeably improved in load balancing and efficient implementation.

IV. FAULT-TOLERANT MECHANISM

The main purpose for designing fault-tolerant is to detect the system fault automatically in real-time, and then take the appropriate controlling or processing strategy, correct mistakes to improve system reliability. The scheduler's Fault-tolerant is able to provide a mechanism, when a resource node goes wrong or failure in the grid environment, it can assigned anew the tasks assigned to that node to other nodes. In other words, once a resource failure in the activation process, it needs redistribute the tasks through scheduling strategy.

A. *Fault Detection*

Due to the characteristics of grid environment, errors may occur at multiple levels. For example, power off, network errors, system crashes, etc. These factors will lead to situations that the resource node is not available, so these errors and failures must be detected and processed timely.

The current grid fault detection mainly links to information services, to detecte the states of resources. This part achieved through the monitoring module, and it is mainly responsible for monitoring registered nodes, detecting node status and polling resource nodes every 5 seconds to test whether it can be used. If the inquiry does not receive any response more than 3 times, consider the node unavailable, and check whether the node has been assigned tasks through the log file, if so, it needs to notify the fault-tolerant module to transfer tasks on that resources node.

B. *Fault-tolerant Processing*

When the fault detection module has detected node's failure, it will start the fault-tolerant module, schedule the tasks to other nodes, and this process is transparent to the users. The most crucial is to choose resources node to receive tasks, that is, re-schedule the task on the fault nodes, map it to the appropriate resources node to run. Because tasks on failure nodes only take a small proportion of the whole tasks submitted, we can use a real-time scheduling strategy to schedule the task to other nodes rapidly, and do not need to re-enter them to the waiting queue which would result in the tasks' waiting in submission queue and reducing the implementation efficiency.

This paper introduces a centralized fault-tolerant load-balancing scheduling strategy to select the appropriate resources node. When monitoring a node fails, use all resources information within the system dynamically for decision-making of the load distribution. Select a node with small load relatively to re-run the task, in order to protect the system load balancing between each node.

1. Centralized load-balancing algorithm

On the basis of the defined factors which affect the load system, the algorithm firstly calculates the value of w which represents the load on each node, calculates the sum and takes the average as the load threshold. Classify the grid nodes into light-load nodes and heavy-duty nodes according to the threshold value and the adjust threshold parameter λ, and then select a light load node to run the job in accordance with the resources information needed by the task.

It is assumed that the grid system has N nodes; the basic steps of the algorithm can be summarized as follows:

① Calculate the current load conditions of N nodes through the formula for load calculating, under the influencing factors and their weights.

② According with the current load condition of each node, calculate the total system load, and the load threshold.

③ Initialize each variable (assume $\lambda=0.1$).

④ According to the classification criteria, classify the grid nodes into light-load nodes and heavy-duty nodes.

⑤ Select a light-load node with the low load in line with the mandate to receive the task.

(1) Load index selection

Any effect of load balancing algorithms is directly dependent on the quality of load estimates and projections, so accurate estimates are necessary. Load indicators commonly used under grid environment include the comprehensive utilization rate of CPU, the length of ready queue, the response time of process, memory size, the comprehensive utilization rate of memory and also the different processing capacity of each heterogeneous node must be considered. Through comprehensive evaluation of these indicators, the load of each node can be obtained.

The comprehensive utilization rate of CPU is defined as time ratio for a certain node to deal with the core process and the user process in a time unit. When it is very low, such as less than default system threshold or the minimum load assessment criteria user defined, we can state that the CPU is idle; when close to 100%,

Table 1 Criterion for qualitative evaluation and load balancing of nodes

State	Load L	Load balancing criterion
Green	L < threshold-λ	Can receive tasks
Yellow	L ≧ threshold – λ and L ≤ threshold +λ	Not to receive tasks is best
Red	L > threshold + λ	Can not accept tasks

measure the load severity of host using CPU queue length. Test shows that run compute-intensive tasks can improve CPU utilization quickly, in a single CPU workstation; a compute-intensive task can make CPU utilization reach 85%-95%. In other words, for scientific computing applications, CPU utilization can only distinguish free state and non-idle state, but it is not able to further distinguish the extent of busy, so it is not sufficient for accurate load decision-making and forecasting. Therefore, we comprehensive survey the load information using weight vector. In this method, the load can be expressed by formula 5-1[4 1]:

$$L = \sqrt{\sum_{i=1}^{n} (k_i \times a_i^2)} \qquad (5\text{-}1)$$

L represents the load value of the local node. a1, a2 ,..., an are the selected load indicators, and k1, k2,..., kn are the weights. The main load indicators pre-designated by users are comparatively large.

(2) Load calculation formula

Set the comprehensive utilization rate of CPU as Yc, the comprehensive utilization rate of memory as Ym, the node capacity as Pa, ready queue length as Rl, among them:

$$\text{Pa}=\frac{\text{Total process number of nodes}}{\text{CPU speed (MIPS)+Total memory(M)}}$$

The parameters such as CPU speed, total memory, the comprehensive utilization rate of CPU and memory, and total process number of nodes can be got through resource discovery module.

Set the parameters for Yc , Ym , Pa and Rl as Wc, Wm, Wpa, Wrl, and Wc＋Wm＋Wpa＋Wrl =1，so the load value L of local node can be calculated by formula 5-1:

$$L=\sqrt{W_c * Y_c^2 + W_m * Y_m^2 + W_{pa} * Pa^2 + W_{rl} * Rl^2}$$

In grid environment, the comprehensive utilization rate of CPU and memory impact the distribution of tasks mostly, so we will set the corresponding parameters Wc and Wm to 0.3, the parameters Wpa and Wrl which corresponds to the node processing power Pa and ready queue length Rl to 0.2.

In order to evaluate the load L qualitatively, we introduce the load adjustment threshold λ, together with the load threshold, as measure standard for L level and load balancing criteria (as shown in Table 1). Its role is to distinguish the light-load nodes more accurately. If we only use the load threshold to classify the node, it will give rise to such a situation: When the threshold is 0.5, then the node whose load value is 0.499 will be classified as a light-load node, and 0.501 as a heavy-duty node. This is likely to give this so-called light-load node an added task. However it can avoid such a situation from happening when we introduce a threshold λ. System load is divided into three states: Green, Red and Yellow, define as follows:

Green status: light load or idle. That means its load value is less than threshold-λ, and it can receive tasks.

Yellow status: the mid-loaded state. That means its load value is between threshold-λ and threshold + λ, and it is best not to receive tasks.

Red state: the heavy-duty state. That means its load value is greater than threshold + λ, and it can not accept the task.

The fault-tolerant scheduling algorithm used in this paper has advantages in simple realization and strong control. When a resource node fails, the algorithm firstly calls the daemon threads to calculate average load value of all available resources, every 5 seconds once, due to the experimental data and traffic volume submitted are relatively small, the load change of each host within 5 seconds is not significant. Short-period (1 second once or more) can reflects the node's load more accurately, but it will increase the entire system cost and burden because of calculating too often. And then, separately divide the nodes whose load is greater and less than the threshold, select a suitable host with light load in the light-load node set, and download the task to perform. This approach not only ensures the task can be re-scheduled quickly after resources failed, but also can balance the load on each node.

V. CONCLUSIONS

The grid resource management system designed in this paper has been applied to the computing grid for seismic data processing, and has achieved good results. The Dmm scheduling strategy shortens the execution time effectively, balances the load of the system, and improves scheduling efficiency. The future work will focus on the research on scheduling occasion, resource description, fault-tolerant processing and other aspects, in order to further optimize and improve the system.

REFERENCES

[1] L.Wang, H.J.Siegel, V.P.Roychowdhury, and A.A.Maciejewski. Task matching and scheduling in heterogeneous computing environments using a genetic-algorithm-based approach[J]. Journal of Parallel and Distributed Computing,47(1):1-15,Nov.1997.

[2] Jian ning Lin, Hui zhong Wu. Scheduling in Grid Computing Environment Based on Genetic Algorithm. Computer Research and Development, 2004;41(12):2196-2199.

[3] Hai Jin, Gang Chen, Mei ping Zhao. Research on a Job Scheduling Model for Fault Tolerant Computational Grid. Computer Research and Development, 2004;41(08):1382-1388.

[4] Krauter K, Buyya R, Maheswaran M. A Taxonomy and Survey of Grid Resource Management Systems. Software Practice and Experience, 2002; 32(2):80-85

[5] Xinjun Chu, Yu qing Fan.The research of PDM based on Web. Journal of Beijing Aeronautics and Astronauts University, 1999(4): 205-207

[6] Ian Foster, Carl Kesselman. Grid: Blueprint for a New Computing Infrastructure.Morgan: Morgan-Kaufman, 1998

[7] Ian Foster, C Kesselman, S Teucke. The Anatomy of the Grid: Enabling Scalable Virtual Organization. International J. Supercomputer Applications, 2001;15 (3):80-84

Zhang Qian

Communication address: College of Computer and Communication Engineering,
China University of Petroleum,
Dongying, Shandong, China, 257062
Telephone: 13780766050
E-Mail: zhangqianupc@163.com

Sampling Theorem Associated with Multiple-parameter Fractional Fourier Transform

Qiwen Ran
National Key Laboratory of Tunable Laser Technology, Harbin Institute of Technology, Harbin, China
Science Research Center, Research Academy of Science and Technology, Harbin Institute of Technology, Harbin, China
Email: qiwenran@hit.edu.cn

Hui Zhao, Guixia Ge, Jing Ma and Liying Tan
National Key Laboratory of Tunable Laser Technology, Harbin Institute of Technology, Harbin, China
Email: {zhaovhit, geguixiahit}@yahoo.com.cn, {majing, tanliying}@hit.edu.cn

Abstract—**We propose a new method for analysis of the sampling and reconstruction conditions of signals by use of the multiple-parameter fractional Fourier transform (MPFRFT). It is shown that the MPFRFT may provide a novel understanding of sampling process. The proposed sampling theorem generalizes classical Shannon sampling theorem and Fourier series expansion, and provides a full-reconstruction procedure of certain signals that are not bandlimited in the conventional Fourier transform domain. An orthogonal basis for the class of signals which are bandlimited in the MPFRFT domain is also given. Experimental results are proposed to verify the accuracy and effectiveness of the obtained results.**

***Index Terms*—sampling theorem, fractional Fourier transform, multiple-parameter fractional Fourier transform**

I. INTRODUCTION

In recent years, the concept of fractional operator has been investigated extensively in many engineering applications and science [1-5]. Fractional operators are defined as fractionalizations of some commonly used operators. In this paper, the fractional Fourier transform (FRFT) are considered. The FRFT, as a generalization of the Fourier transform, has different kinds of mathematical definitions [6-9]. This fact enables us to represent signals in different ways. Shih [10] proposed a method to fractionalize the Fourier transform as a composition of the given signal, its ordinary Fourier transform and their reflected versions, only according to three postulates that the FRFT should obey. We generalize the weighted coefficients of the FRFT proposed by Shih to contain two vector parameters. Therefore a generalized FRFT is defined by replacing the weighted coefficients with the generalized ones, which is regarded as the multiple-parameter fractional Fourier transforms (MPFRFT).

Sampling theorem plays a crucial role in signal processing and communications [11-13]. In the sampling problem, the objective is to reconstruct a signal from its samples. For a bandlimited signal, Shannon sampling theorem provides a full reconstruction by its uniform samples with a sampling rate higher than its Nyquist frequency [11]. For non-bandlimited signals, several sampling criteria have been proposed associated with wavelet transform and Wigner distribution function etc. [14-19]. Herein we propose another transform for investigating sampling: the MPFRFT. We show that the MPFRFT may provide additional insights that are not observed with traditional Fourier transform, wavelet transform etc.

The main result from our MPFRFT-based sampling analysis is a generalization of Shannon sampling theorem and Fourier series expansion. The proposed sampling theorems enable us to sample and reconstruct certain signals that are not bandlimited in the conventional Fourier transform domain. In section II, the definitions of one dimensional (1D) and two dimensional (2D) MPFRFT are defined. The sampling analysis based on MPFRFT is given in Section III. An orthogonal basis for the class of bandlimited signals in MPFRFT domain is also given in section III. In section IV, experimental results are proposed to demonstrate the effectiveness of the proposed sampling theorems. Section V concludes this paper.

II. MULTIPLE-PARAMETER FRACTIONAL FOURIER TRANSFORM

Let Θ be an operator. Θ : $\Theta\left[g(x)\right]=G(u)$. It is generally agreed that the fractional operation Θ^{α} of operation Θ should satisfy the following postulates:

i. Continuity postulate: Θ^{α} should be continuous for all real values α .

ii. Boundary postulate:

$$\Theta^{0}\left[g(x)\right]=g(x), \Theta^{1}\left[g(x)\right]=G(u) \tag{1}$$

iii. Additivity postulate:

$$\Theta^{\beta}\left\{\Theta^{\alpha}\left[g(x)\right]\right\}=\Theta^{\alpha}\left\{\Theta^{\beta}\left[g(x)\right]\right\}=\Theta^{\alpha+\beta}\left[g(x)\right] \tag{2}$$

According to above postulates, one can fractionalize any operation in different ways.

A. One dimensional MPFRFT

It is well known that the Fourier transform F is periodic with periodicity 4. Therefore, it is reasonable to

doi:10.4304/jcp.5.5.695-702

assume that any fractional operator F^{α} of the Fourier transform F is a weighted combination of the four basic operators F^0 , F^1 , F^2 and F^3 . Analogous to Shih's technique [10], we can define the fractional Fourier transform F^{α} with order α as

$$F^{\alpha}\left(x\right)=\sum_{k=0}^{3}p_k\left(\alpha\right)F^k\left(x\right) \tag{3}$$

where the weighted coefficients are the functions of transform order α. According to above three postulates, the coefficients should satisfy the following conditions

i. The coefficients are continuous functions of transform order α;

ii. When α is an integer, the coefficients should be certain values which serve as boundary conditions, see Table I;

iii. The coefficients should satisfy the following coupling equations

$$\begin{aligned}
p_0(\alpha+\beta)&=p_0(\alpha)p_0(\beta)+p_1(\alpha)p_3(\beta)+p_2(\alpha)p_2(\beta)+p_3(\alpha)p_1(\beta)\\
p_1(\alpha+\beta)&=p_0(\alpha)p_1(\beta)+p_1(\alpha)p_0(\beta)+p_2(\alpha)p_3(\beta)+p_3(\alpha)p_2(\beta)\\
p_2(\alpha+\beta)&=p_0(\alpha)p_2(\beta)+p_1(\alpha)p_1(\beta)+p_2(\alpha)p_0(\beta)+p_3(\alpha)p_3(\beta)\\
p_3(\alpha+\beta)&=p_0(\alpha)p_3(\beta)+p_1(\alpha)p_2(\beta)+p_2(\alpha)p_1(\beta)+p_3(\alpha)p_0(\beta)
\end{aligned} \tag{4}$$

In order to calculate the values of the coefficients, we take the following transformation

$$\begin{pmatrix}q_0(\alpha)\\q_1(\alpha)\\q_2(\alpha)\\q_3(\alpha)\end{pmatrix}=\begin{pmatrix}1&1&1&1\\1&-i&-1&i\\1&-1&1&-1\\1&i&-1&-i\end{pmatrix}\begin{pmatrix}p_0(\alpha)\\p_1(\alpha)\\p_2(\alpha)\\p_3(\alpha)\end{pmatrix} \tag{5}$$

Using this transformation, we get the following four equations with much simpler forms

$$\begin{pmatrix}q_0(\alpha+\beta)\\q_1(\alpha+\beta)\\q_2(\alpha+\beta)\\q_3(\alpha+\beta)\end{pmatrix}=\begin{pmatrix}q_0(\alpha)q_0(\beta)\\q_1(\alpha)q_1(\beta)\\q_2(\alpha)q_2(\beta)\\q_3(\alpha)q_3(\beta)\end{pmatrix} \tag{6}$$

The solutions of (6) are simple exponential functions. Considering the new boundary conditions for $q_k(\alpha)$ k =0, 1, 2, 3 shown in Table I, we can get the solution of (6) as

$$q_k(\alpha)=\exp\{-2\pi i\alpha(4m_k+1)(4n_k+k)/4\} \tag{7}$$

where parameter vectors

$$m=\left(m_0,m_1,m_2,m_3\right)\in Z^4,\, n=\left(n_0,n_1,n_2,n_3\right)\in Z^4$$

are two arbitrary 4-dimensional integer vectors. Taking the inverse transform of (5), we can obtain the weighted coefficients $p_k(\alpha), k=0,1,2,3$ as

$$p_k(\alpha)=\frac{1}{4}\sum_{l=0}^{3}\exp\{-2\pi i/4[(4m_l+1)\alpha(4n_l+l)-kl]\} \tag{8}$$

Thus the fractional Fourier transform with order α of signal f , or briefly, the (α)-FRFT of f , can be defined as

$$F^{\alpha}\left(x\right)=\sum_{k=0}^{3}p_k\left(\alpha,m,n\right)F^k\left(x\right) \tag{9}$$

with the weighted coefficients $p_k\left(\alpha,m,n\right)$ given by (8). Due to the additional freedom degrees provided by parameter vectors $m,n,$ we call this kind of FRFT multiple-parameter fractional Fourier transform (MPFRFT).

Note that when $m=n=(0,0,0,0)$, the MPFRFT reduces to the FRFT proposed by Shih. As shown in Fig. 1, the randomicity of parameter vectors m and n provides us more choices to represent signals.

B. two dimensional MPFRFT

Now we extend the 1D MPFRFT to 2D case. Similar to the 1D case, the 2D (α) -MPFRFT of signal $f(x,y)$ can be defined as

$$F^{\alpha}[f(x,y)]=\sum_{k=0}^{3}p_k\left(\alpha,m,n\right)F^k[f(x,y)] \tag{10}$$

where F^k denotes the k -order 2D Fourier transform, the weighted coefficients $p_k\left(\alpha,m,n\right)$, $k=0,1,2,3$ are the same as (8).

III. SAMPLING THEOREM ASSOCIATED WITH MPFRFT

A. Sampling theorem associated with one dimensional MPFRFT

A signal f is said to be σ bandlimited in (α)-MPFRFT domain, if there exists a positive σ such that

$$F^{\alpha}[f](u)=0,|u|>\sigma\,.$$

Theorem 1: Suppose $f(t)$ is σ bandlimited in (α)-MPFRFT domain, then $f(t)$ can be uniquely determined by the samples of its ($\alpha-1$)-MPFRFT, and can be completely reconstructed by the following sampling formula:

$$f(t)=\sum_{n}F^{\alpha-1}[f](t_n)\phi(t,t_n) \tag{11}$$

where $t_n\leq n/(2\sigma)$, $\phi(t,t_n)$ is the ($1-\alpha$)-MPFRFT of $\mathrm{sinc}[2\sigma(t-t_n)]$ and $\mathrm{sinc}(t)=\sin(\pi t)/(\pi t)$.

TABLE I.
BOUNDARY VALUES FOR THE COEFFICIENTS $p_k(\alpha)$ AND $q_k(\alpha)$.

α	$p_0(\alpha)$	$p_1(\alpha)$	$p_2(\alpha)$	$p_3(\alpha)$	$q_0(\alpha)$	$q_1(\alpha)$	$q_2(\alpha)$	$q_3(\alpha)$
0	1	0	0	0	1	1	1	1
1	0	1	0	0	1	$-i$	-1	i
2	0	0	1	0	1	-1	1	-1
3	0	0	0	1	1	i	-1	$-i$

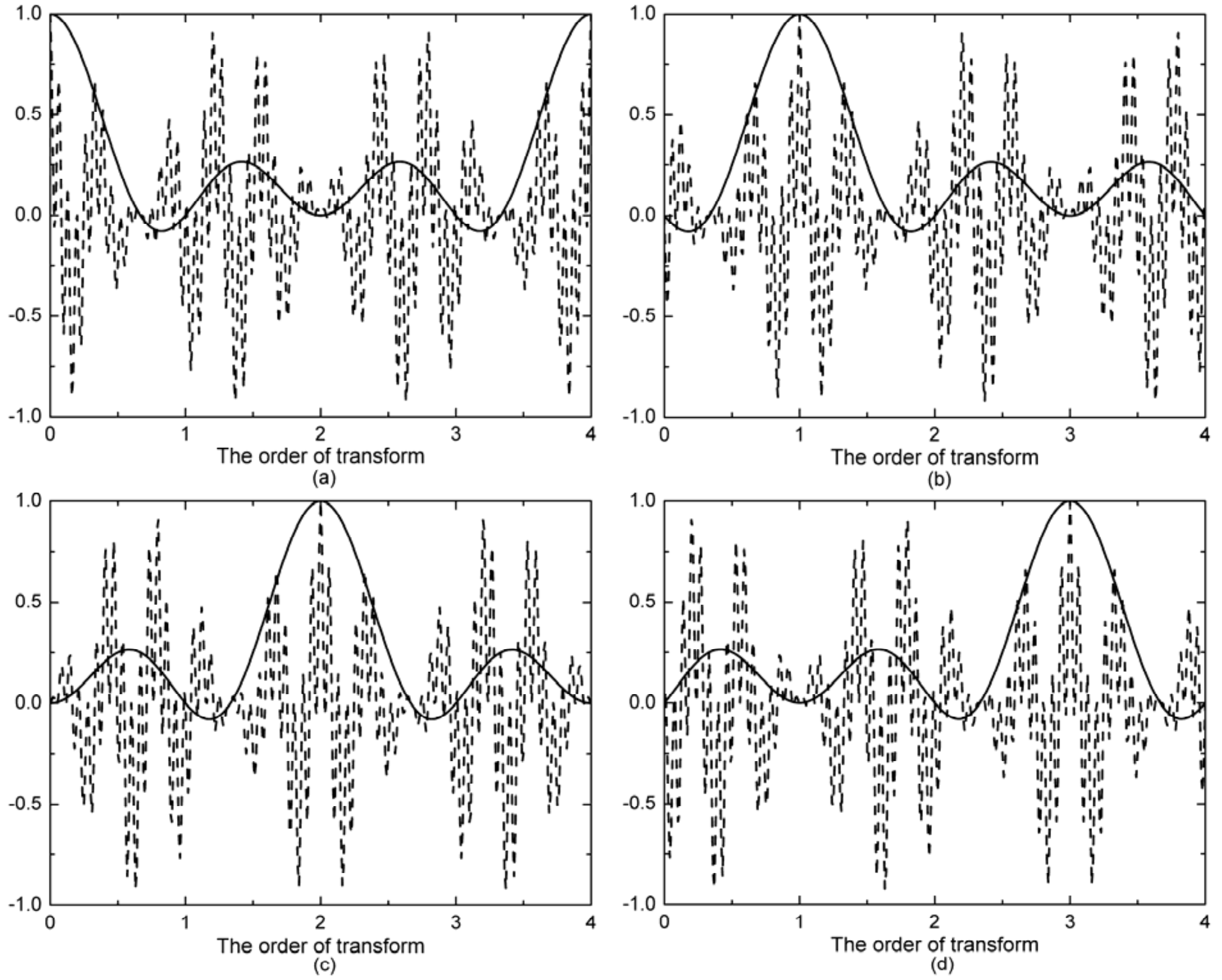


Figure 1. Real parts of the weighted coefficients, plotted as a function of the transform order, with different parameter vectors: Solid curves with parameter vectors $m = n = (0,0,0,0)$ and dashed curves with parameter vectors $m = (1,3,0,2)$ and $n = (3,0,2,1)$. (a) $p_0(\alpha, m, n)$, (b) $p_1(\alpha, m, n)$, (c) $p_2(\alpha, m, n)$, (d) $p_3(\alpha, m, n)$.

Proof: Let $g(u)$ denote the $(\alpha - 1)$-MPFRFT of $f(t)$. It is a common engineering practice to model the sampling process by a multiplication with a sampling sequence of

$$\frac{1}{u_s}\operatorname{comb}(\frac{u}{u_s})$$

where u_s is the sampling period. Without loss of generality, we let $u_s > 0$. Here the corresponding sampled signal representation is

$$g_s(u) = g(u)\frac{1}{u_s}\operatorname{comb}(\frac{u}{u_s}) \tag{12}$$

with

$$\operatorname{comb}(\frac{u}{u_s}){=}u_s\sum_n \delta(u - u_s)$$

It is known that the Fourier transform of the multiplication of two signals corresponds to a convolution of the Fourier transforms of each signal. Taking into account the properties of the comb function and the additive property of MPFRFT, we can write the Fourier transform $G_s(w)$ of $g_s(u)$ as

$$\begin{aligned} G_s(w) &= F^{\alpha}[f](w) * \operatorname{comb}(u_s w) \\ &= \frac{1}{u_s}\sum_n F^{\alpha}[f](w - \frac{n}{u_s}) \end{aligned} \tag{13}$$

where superscript $*$ denotes a convolution. Thus the sampling process results in a periodization of the (α)-MPFRFT of f, as illustrated in Fig. 2. Choose ideal low-pass filter

$$\mathrm{R}(w){=}\begin{cases} 1, |w| \le \sigma \\ 0, |w| > \sigma \end{cases} \tag{14}$$

and $u_s = 1/(2\sigma)$, we have

$$G_s(w)R(w) = 2\sigma F^{\alpha}[f](w) \tag{15}$$

By the additive property of MPFRFT, in the $(\alpha - 1)$ MPFRFT domain, we have

$$\begin{aligned} g(u) &= g_s(u) * \operatorname{sinc}(2\sigma u) \\ &= \sum_n g(t_n)\operatorname{sinc}[2\sigma(u - t_n)] \end{aligned} \tag{16}$$

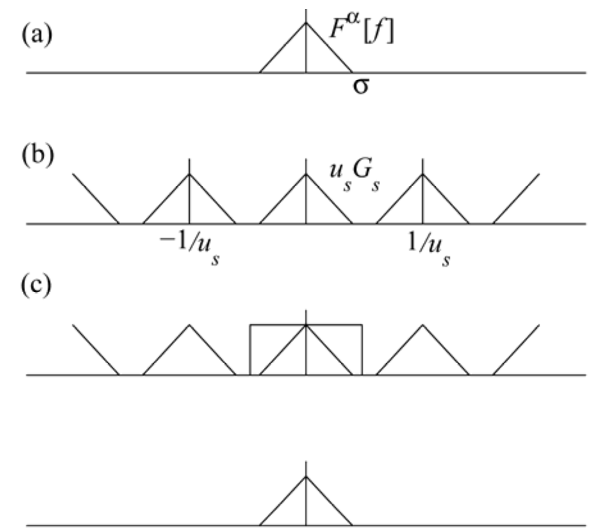


Figure 2. Effect of sampling in $(\alpha-1)$ -MPFRFT domain. (a) The (α) -MPFRFT of analog input signal f . (b) Sampling process results in a periodization of the (α) -MPFRFT. (c) The analog signal is reconstructed by ideal low-pass filtering.

Since $f(t)$ is the $(1-\alpha)$ -MPFRFT of $g(u)$, we have

$$\begin{aligned} f(t) &= F^{1-\alpha}[g(u)](t) \\ &= \sum_n g(t_n) F^{1-\alpha}\left\{\mathrm{sinc}[2\sigma(u\text{-}t_n)]\right\}(t) \end{aligned} \quad (17)$$

Obviously, when $\alpha=1$ and $m=n=(0,0,0,0)$, Eq. (11) reduces to

$$f(t) = \sum_n f(t_n)\,\mathrm{sinc}[2\sigma(t-t_n)] \quad (18)$$

which means that Theorem 1 includes Shannon sampling theorems as special case. Compared with Shannon sampling theorem, sampling condition is enlarged and the samples are not limited to time domain.

It is interesting to see that when $\alpha=0$ and $m=n=(0,0,0,0)$, Eq. (11) reduces to

$$f(t) = \begin{cases} (2\sigma)^{-1}\sum_n F(-t_n)e^{-i2\pi t_n t}, & |t|\le\sigma \\ 0, & |t|>\sigma \end{cases} \quad (19)$$

which means that a timelimited signal can be represented by a Fourier series. Thus, classical Fourier series expansion can also be viewed as a sampling formula in the sense of MPFRFT. Furthermore, the proposed sampling theorem based on MPFRFT gives a continuous conversion from Fourier series expansion to Shannon sampling theorem when $m=n=(0,0,0,0)$ and transform order α ranges from 0 to 1.

Theorem 2: Let H denote the class of signals which are σ bandlimited in (α)-MPFRFT, then we have the following results:

1) The sequence $\{\phi(t,t_n)\}$ forms an orthogonal basis for H ;

2) With respect to above basis, the coordinates of signal are actually the uniform samples of its $(\alpha-1)$-MPFRFT.

Proof: 1) Let H_1 denote the class of signals which are σ bandlimited in the conventional Fourier transform domain. It has been given before that $\{\phi(t,t_n)\}$ is the $(1-\alpha)$-MPFRFT of $\mathrm{sinc}[2\sigma(t-t_n)]$. Since the MPFRFT is a unitary mapping of $L^2(R)$ into itself under which H is the image of H_1, it follows that $\mathrm{sinc}[2\sigma(t-t_n)]$ is an orthogonal basis for H_1 if and only if $\{\phi(t,t_n)\}$ is an orthogonal basis for H . It is well known that $\mathrm{sinc}[2\sigma(t-t_n)]$ is an orthogonal basis for H_1. Therefore, $\{\phi(t,t_n)\}$ forms an orthogonal basis for H .

2) From 1) for any $f(t)\in H$, we have

$$f(t) = \sum_n c_n \phi(t,t_n) \quad (20)$$

where $c_n, n\in Z$ denote the coordinates of $f(t)$ and can be calculated as

$$c_n = \left\langle f(t), \frac{\phi(t,t_n)}{\|\phi(t,t_n)\|^2} \right\rangle \quad (21)$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product and $\|\cdot\|$ is the 2-norm. By the additive property of MPFRFT, it is easy to see that the (α)-MPFRFT $\psi(w)$ of $\{\phi(t,t_n)\}$ is actually the Fourier transform of $\mathrm{sinc}[2\sigma(u-t_n)]$. Thus $\psi(w)$ equals to $(2\sigma)^{-1}e^{-i2\pi t_n t}$ for $|t|\le\sigma$ and equals to zero otherwise. The unitary property of MPFRFT yields that

$$\|\phi(t,t_n)\|^2 = \|\psi(w)\|^2 = \frac{1}{2\sigma} \quad (22)$$

So

$$\begin{aligned} c_n &= 2\sigma\langle f(t), \phi(t,t_n)\rangle \\ &= 2\sigma\left\langle F^{\alpha}[f(t)](w), F^{\alpha}[\phi(t,t_n)](w)\right\rangle \\ &= \int_{-\infty}^{\infty} F^{\alpha}[f](w)e^{i2\pi t_n w}dw \\ &= F^{\alpha-1}[f](t_n) \end{aligned} \quad (23)$$

It can be seen from Theorem 2 that $\{\phi(t,t_n)\ , n\in Z\}$ and their any linear combinations are σ bandlimited in (α) -MPFRFT domain, and can then be completely reconstructed according to sampling formula (11). Obviously, when $\alpha\ne 1$, $\{\phi(t,t_n)\ , n\in Z\}$ are not bandlimited in the conventional Fourier transform domain. Therefore, Theorem 1 provides a full-reconstruction of certain signals that are not bandlimited in the conventional Fourier transform domain.

B. Sampling theorem associated with two dimensional MPFRFT

For notational simplicity, 2D variables (x,y) are denoted as a vector $X=[x\ \ y]^{\mathrm{T}}$. Let $W(X)$ denote the $(\alpha-1)$-2D MPFRFT of f . Let us denote the sampled version of $W(X)$ by $W_s(X)$ for which the periodic sampling geometry is indicated by the sampling matrix V as

$$W_s(X) = W(X)\sum_N \delta(X - VN) \qquad (24)$$

where $\delta(X)$ is the 2D impulse function, $N = [n \;\; m]^T$ and n and m are integers. Sampling matrix

$$V = \begin{bmatrix} x_s & 0 \\ 0 & y_s \end{bmatrix}$$

with x_s and y_s are the distances between samples in the x and y directions, respectively. In the Fourier domain, multiplication corresponds to a convolution. Therefore, by the additive property of MPFRFT, the Fourier transform $\overline{W}_s$ of W_s can be written as

$$\begin{aligned} \overline{W}_s(U) &= \mathcal{F}[W_s(X)] \\ &= \mathcal{F}^{\alpha}[f](U) ** \frac{1}{|\det V|}\sum_N \delta(U - V^{-1}N) \\ &= \frac{1}{|\det V|}\sum_N \mathcal{F}^{\alpha}[f](U - V^{-1}N) \qquad (25) \end{aligned}$$

where $U = [u \;\; v]^T$, as usual. The double asterisks denotes 2D convolution operator. It can be seen that the Fourier transform of W_s is formed from infinite superposed, shifted replicas of the (α)-MPFRFT of the original signal $f(X)$. The effect of sampling in the $(\alpha - 1)$-MPFRFT domain is illustrated in Fig. 3.

As shown in Fig. 3, the sampling process results in superposed, shifted replicas of the (α)-MPFRFT of the original signal $f(X)$ and the replicas are located at $V^{-1}N$. Therefore, if $f(X)$ is bandlimited in the (α)-MPFRFT domain, say within a band $U \in B =$ $[-b_u, b_u] \times [-b_v, b_v]$, and if the sampling matrix V is chosen to satisfy nonoverlapping replicas in the (α)-MPFRFT domain, say $x_s = (2b_u)^{-1}$ and $y_s = (2b_v)^{-1}$, then $\mathcal{F}^{\alpha}[f](U)$ can be fully recovered by low-pass filtering the Fourier transform $\overline{W}_s$ of W_s . Mathematically, we have

$$\overline{W}_s(U)R(U) = \frac{1}{|\det V|}\mathcal{F}^{\alpha}[f](U) \qquad (26)$$

where

$$R(U) = \begin{cases} 1, & U \in B \\ 0, & \text{else} \end{cases}$$

Therefore, $W(X)$ can be written as

$$\begin{aligned} W(X) &= W_s(X) ** \operatorname{sinc}(V^{-1}X) \\ &= \sum_N W(VN)\operatorname{sinc}[V^{-1}(X - VN)] \qquad (27) \end{aligned}$$

Since f is the $(1-\alpha)$-MPFRFT of $W(X)$, we thus have

$$f(X) = \sum_N W(VN)\gamma_n(X) \qquad (28)$$

where $\gamma_n(X)$ is the $(1-\alpha)$ -MPFRFT of $\operatorname{sinc}[V^{-1}(X - VN)]$.

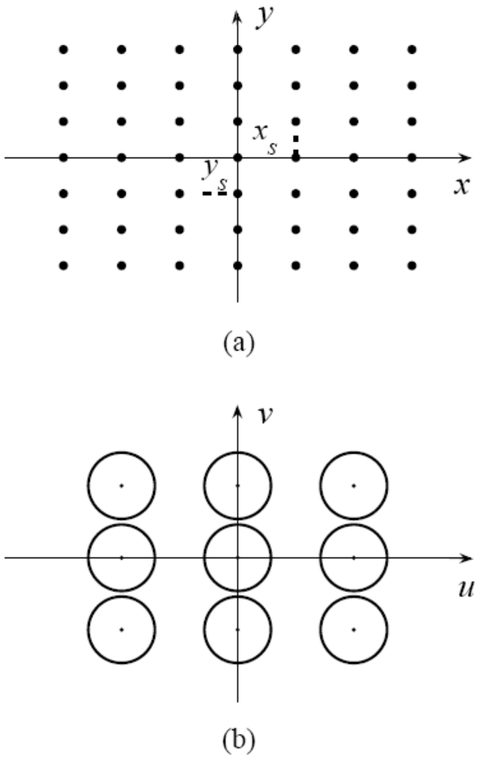


Figure 3. Effect of sampling in the $(\alpha - 1)$ – 2D MPFRFT domain. (a) Sampling lattice (spaced by x_s and y_s in the x and y directions, respectively) of the $(\alpha - 1)$ -MPFRFT domain, (b) The sampling process results in replications of the (α) -MPFRFT.

The above discussion yields the following sampling theorem.

Theorem 3: Suppose $f(X)$ is $B = [-b_u, b_u] \times [-b_v, b_v]$ bandlimited in (α) -2D MPFRFT domain, then $f(X)$ can be fully reconstructed from its $(\alpha - 1)$-2D MPFRFT domain samples according to formula (28).

IV. Simulation Examples

Figures 4 and 5 depict the results of a numerical experiment demonstrating the effectiveness of sampling formula (11) presented in this work. Figs. 4(a) and (b) show a MPFRFT pair with $\alpha = 0.2$, $m = (1,3,0,2)$ and $n = (3,0,2,1)$. The MPFRFTed signal $F^{\alpha}[f](w)$ equals to e^{iw}, for $w \in$ [-1,1], and equals to zero otherwise; that is the original signal $f(t)$ illustrated in Fig. 4(a) is $\sigma = 1$ bandlimited in (0.2)-MPFRFT domain with parameter vectors $m = (1,3,0,2)$ and $n = (3,0,2,1)$. Therefore, from Theorem 1, we can reconstruct $f(t)$ by use of the samples of it (-0.8)-MPFRFT and interpolation functions $\{\phi(t, t_n)\}$. The (-0.8)-MPFRFT domain samples c_n and the reconstructed signal based on Theorem 1 are plotted in Figs. 4(c) and (d), respectively. Several interpolation functions $\{\phi(t, t_n)\}$ in (11) are illustrated in Fig. 5. We can see that an almost perfect reconstruction of $f(t)$ can be obtained.

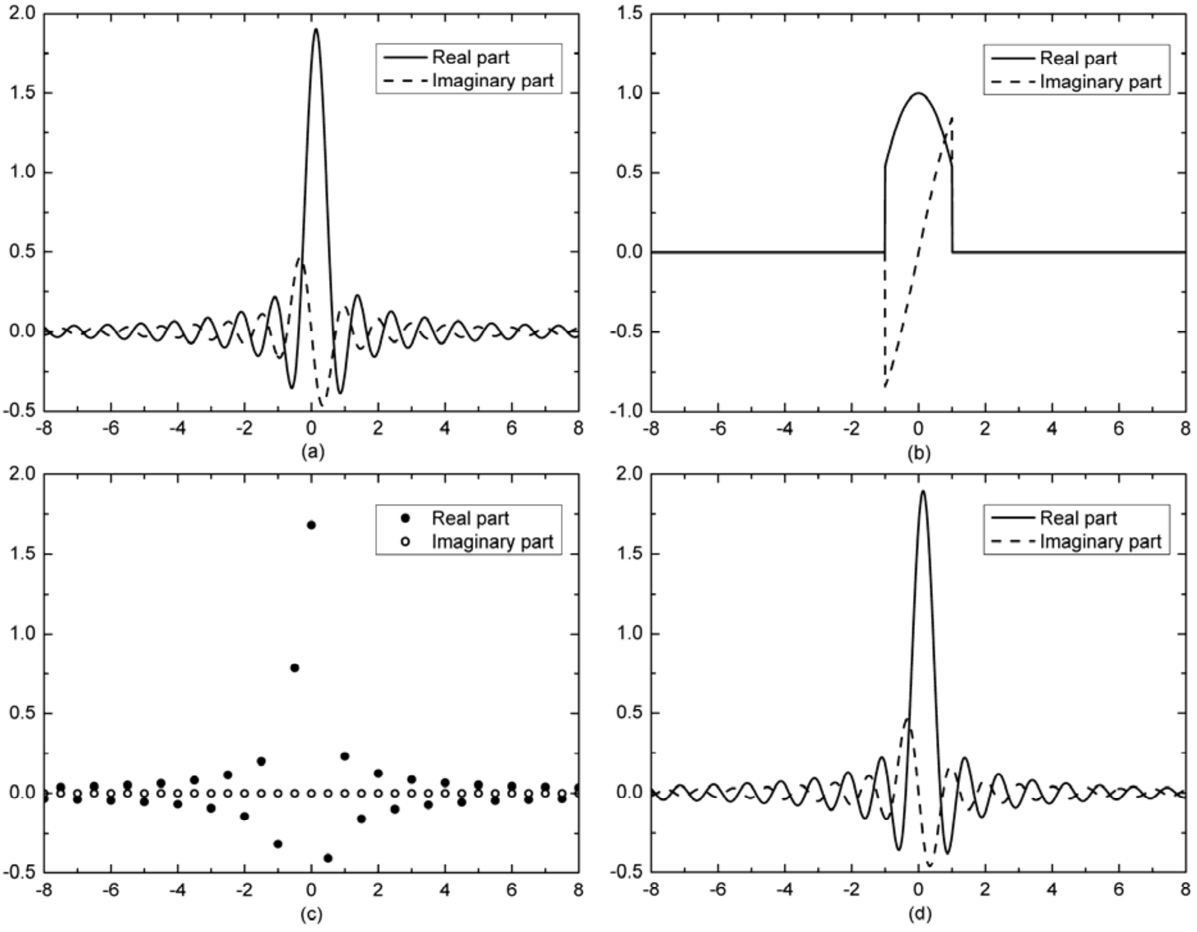


Figure 4. Example of reconstruction of a signal based on Theorem 1. (a) The original signal $f(t)$. (b) The (0.2)-MPFRFT $F^{\alpha}[f](w)$ of $f(t)$ with $m=(1,3,0,2)$, $n=(3,0,2,1)$. (c) The (-0.8)-MPFRFT domain samples c_n . (d) Reconstructed signal based on Theorem 1. It can be seen that an almost perfect reconstruction is obtained.

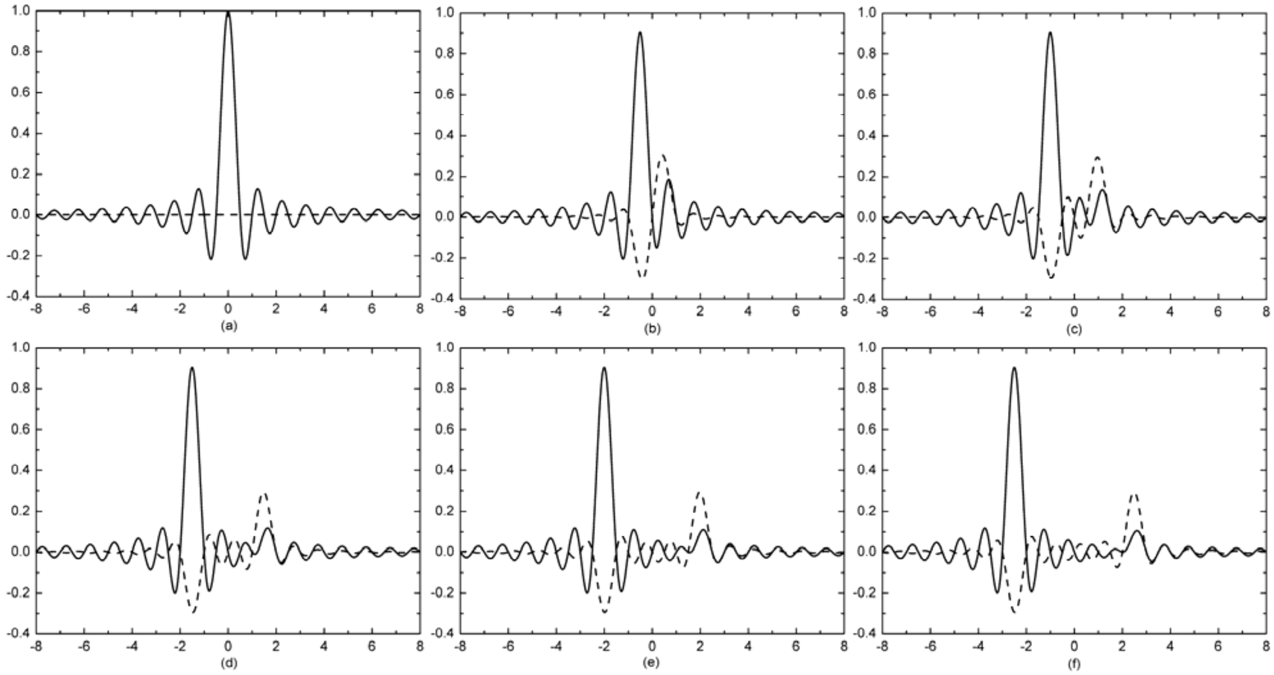


Figure 5. Interpolation functions $\{\phi(t,t_n)\}$ with $\alpha=0.2$ $\sigma=1$, and (a) n = 0, (b) n = 1, (c) n = 2, (d) n = 3, (e) n = 4 , (f) n = 5. Solid curves stand for real parts and dashed curves stand for imaginary parts.

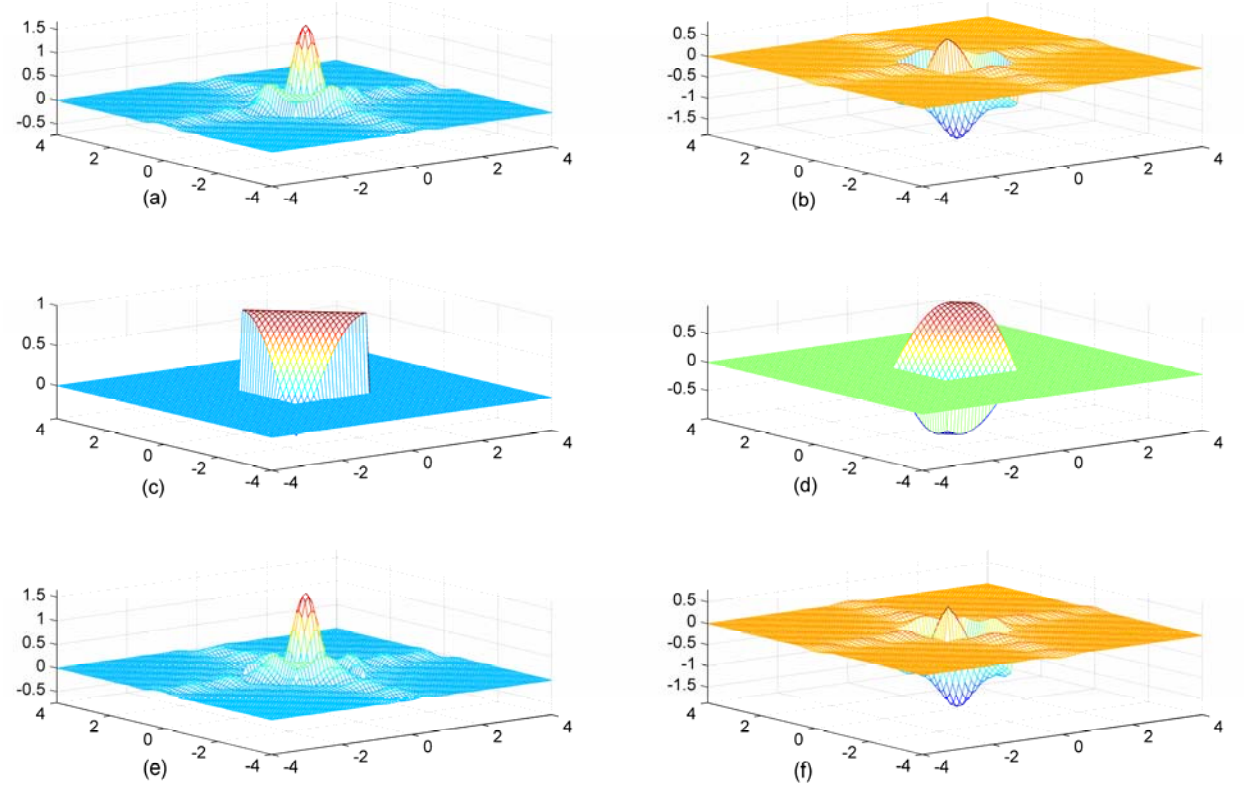


Figure 6. Reconstruction of a 2D signal based on Theorem 3. (a) Real part of the original signal $f(x,y)$, (b) Imaginary part of f, (c) Real part of the (0.6)-MPFRFT $F^{0.6}f$ with parameter vectors $m=(2,5,3,1)$ and $n=(7,4,1,2)$ of f, (d) Imaginary part of $F^{0.6}f$. (e) Real part of the reconstruction signal, and (f) Imaginary part of the reconstruction signal. It can be seen that an almost full reconstruction is obtained.

Simulation example shown in Fig. 6 gives further insight. The real and imaginary parts of a 2D signal are shown in Figs. 6(a) and (b), respectively. The real and imaginary parts of its (0.6)-MPFRFT with parameter vectors $m=(2,5,3,1)$ and $n=(7,4,1,2)$ are shown in Figs. 6(c) and (d), respectively. It can be seen from (c) and (d) that the original signal is $B=[-1,1]\times[-1,1]$ bandlimited in (0.6)-MPFRFT with parameter vectors $m=(2,5,3,1)$ and $n=(7,4,1,2)$. Therefore, from Theorem 3, the original signal can be fully reconstructed by the samples of its (-0.4)-MPFRFT. The (-0.4)-MPFRFT is sampled with a sampling matrix V,

$$V=\begin{bmatrix}0.5 & 0\\ 0 & 0.5\end{bmatrix}$$

Figs. 6(e) and (f) display the real and imaginary parts, respectively, of the reconstruction signal. It can be seen that an almost perfect reconstruction is obtained.

V. Conclusion

In this paper, we first generalized the fractional Fourier transform (FRFT) proposed by Shih to multiple-parameter FRFT (MPFRFT) and extend the 1D MPFRFT to 2D case. Then we proposed a new method for analysis of sampling and reconstruction of signals by use of the MPFRFT. On the basis of observations in MPFRFT domain, we derived a generalization of Shannon sampling theorem and Fourier series extension. The proposed theorem unifies classical Shannon sampling theorem with Fourier series expansion. It also provides a full-reconstruction procedure of certain signals that are not band-limited in the conventional Fourier domain. An orthogonal basis for the class of bandlimited signals in MPFRFT domain is also given, with respect to which the coordinates of signal are actually the samples of its $(\alpha-1)$-MPFRFT. Experimental results have verified the accuracy and effectiveness of the obtained results.

Acknowledgment

This research is funded by the National Natural Science Foundation of China (NSFC) (No. 10374023 and 60432040) and the National Basic Research Program of China (No. 2007CB310606).

References

[1] S.C. Pei and J.J. Ding, "Relations between fractional operations and time-frequency distributions, and their applications", *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1638-1655, 2001.

[2] O. Akay and G.F. Boudreaux-Bartels, "Fractional convolution and correlation via operator methods and an application to detection of linear FM signals", *IEEE Trans. Signal Process.*, vol. 49, no. 5, pp. 979-993, 2001.

[3] O. Akay and G.F. Boudreaux-Bartels, "Unitary and hermitian fractional operators and their relation to the

fractional Fourier transform," *IEEE Signal Process. Lett,* vol. 5, no. 12, pp. 312-314, 1998.

[4] G. Cariolaro, T. Erseghe and P. Kraniauskas, "The fractional discrete cosine transform", *IEEE Trans. Signal Process.*, vol. 50, no. 4, pp. 902-911, 2002.

[5] S.C. Pei and J.J. Ding, "Fractional cosine, sine, and Hartley transforms," *IEEE Trans. Signal Process.*, vol. 50, no. 7, pp. 1661-1680, Jul. 2002.

[6] G.Cariolaro, T. Erseghe, P. Kraniauskas and N. Laurenti, "Multiplicity of fractional Fourier transforms and their relationships," *IEEE Trans. Signal Process.*, vol. 48, pp. 227-241, Jan. 2000.

[7] G. Cariolaro, T. Erseghe, P. Kraniauskas and N. Laurenti. "A unified framework for the fractional Fourier transform", *IEEE Trans. Signal Process.*, vol. 16, no. 12, pp. 3206-3219, 1998.

[8] Q.W. Ran, D.S. Yeung, C.C. Tsang and Q. Wang, "General multifractional fourier transform method based on the generalized permutation matrix group," *IEEE Trans. Signal Process.*, vol. 53, pp. 83-98, 2005.

[9] T. Erseghe, P. Kraniauskas and G. Cariolaro. "Unified fractional Fourier transform and sampling theorem," *IEEE Trans. Signal Process,* vol. 47, pp. 3419-3423, 1999.

[10] C.C. Shih, "Fractionalization of fourier transform," *Opt. Commun.*, vol. 118, pp. 495-498, Aug. 1995.

[11] A.J. Jerri, "The shannon sampling theorem-its various extensions and applications: A tutorial review," *Proc. IEEE*, vol. 65, pp. 1565-1596, Nov. 1977.

[12] A. Jerri, "On the equivalence of Kramer's and Shannon's sampling theorems," *IEEE Trans. Inform. Theory,* vol. 15, no. 4, pp. 497-499, Jul. 1969.

[13] M. Unser, "Sampling-50 years after shannon," *Proc. IEEE*, vol. 88, pp. 569-587, 2000.

[14] P.L. Butzer and R.L. Stens, "Sampling theory for not-necessarily band-limited functions: A historical overview," *SIAM Rev.*, vol. 34, pp. 40-53, 1992

[15] L. Onural, "Sampling of the diffraction field," *Appl. Opt.*, vol. 39, pp. 5929-5935, 2000.

[16] A. Stern and B. Javidi, "Sampling in the light of wigner distribution," *J. Opt. Soc. Am. A,* vol. 21, pp. 360-366, 2004.

[17] C. Zhao and P. Zhao, "Sampling theorem and irregular sampling theorem for multiwavelet subspaces," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 705-713, Feb. 2005.

[18] M. Unser, "A generalized sampling without bandlimiting constraints," *IEEE Trans. Circuits Syst.*, vol. 45, pp. 959-969, Aug. 1998.

[19] X.G. Xia and Z. Zhang, "On sampling theorem, wavelets, and wavelet transforms," *IEEE Trans. Signal Process.*, vol. 41, pp. 3524-3534, 1993.

Qiwen Ran is a professor at Harbin Institute of Technology. He received the Ph.D degree in computer applications from Harbin Institute of Technology in 1999. His research interests are wavelets theory, fractional Fourier transforms and computer applications.

Hui Zhao is a Ph.D. candidate at Harbin Institute of Technology. Her research interests are fractional Fourier transform, linear canonical transform and signal processing.

Guixia Ge received the Master degree in applied mathematics from Harbin Institute of Technology in 2009. Her research interests are wavelets theory and fractional Fourier transforms.

Jing Ma is a professor at Harbin Institute of Technology. He received the Ph.D degree in physical electronics from Harbin Institute of Technology in 2001. His research interests are satellite optical communication and wavelet optics information processing.

Liying Tan is a professor at Harbin Institute of Technology. She received the Ph.D degree in physical electronics from Harbin Institute of Technology in 2004. Her research interests are optical information processing and satellite optical communication.

The Optimized Comparison of the Gray Model Improved by Posterior-Error-Test and SVM Modified by Markov Residual Error in the Long-medium Power Load Forecast

Li wei
Department of Economics and Management, North China Electric Power University, Baoding, Hebei, 071003, China
hd11111@126.com

Zhang zhen-gang
Department of Economics and Management, North China Electric Power University, Baoding, Hebei, 071003, China
zhangzhengangwf@163.com

Abstract — **Generally, the long-medium power load forecasting sequence has small sample, stochastic growth and nonlinear wave characteristics. Gray and SVM model could reflect the relationship between growing characters--tics and nonlinear characteristics to the series effectively and make fitting calculation. The paper modifies the proposed gray model through posterior-error-test and compares the predictive value of power load when the evaluation result is best with the optimal result forecast by SVM that is modified by Markov residual. Then we can find which model is the better. As result, we can see that Markov could well reflect randomness that produced by the system involve with many complex factors. A forecast model based on SVM algorithm is established, the series of historical load variables is rolling forecasted. It is proved that the presented forecast method is superior obviously to traditional methods through empirical study, and it can be used generally.**

Index Terms —**Posterior-error-test, GM(1,1) model, Markov, SVM residual error, power load forecast**

I. Introduction

Long-medium power load forecast is the necessary condition and foundation for the program, design, researching, production and operation of power system, the accurate result of the load forecast has an direct influence on the capital management of power construction. Especially, the power supply and consumption which are lead by various undetermined factors is surge. The power load will have excellent effects.

The gray SVM method mentioned in [4][5] gained favorable results. Using the combined weight reveal the future development trend of system. So the residual error modified method was used in power load forecasting. The credibility and precision can be further improved in practice, but lack more modification to errors. The posterior-error-test and weighted Markov residual error method [6] is presented to modify forecast results, and it has a good effect. According to an appropriate normalization method, we consider the alterative value of historical load as input variables for GM(1.1) or SVM method, and get GM(1.1) or SVM load forecast value by the way of intelligent optimization forecast. At last, we obtain the final forecast results through the improved Markov process to modify the forecast error, the final results are tested by empirically and an ideal result.

II. GM(1,1) Model

GM (1, 1) model is one of the common gray models, it consists by the single variable differential equation, and it is the effective model for power load forecasting.

Supposing there is a data list named $x^{(0)}$

$$x^{(0)} = \left[x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n)\right] \tag{1}$$

Using 1-AGO to create the progression list

$$x^{(1)} = \left[x^{(1)}(1), x^{(1)}(2), \cdots, x^{(1)}(n)\right]$$

$$x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i)\left(k = 1, 2, \cdots, n\right) \tag{2}$$

$x^{(1)}$ According to the model as follows

$$\frac{dx^{(1)}}{dt} + ax^{(1)} = u \tag{3}$$

Using the least square method to make the parameter $\hat{a}, \hat{u}$

$$\hat{A} = (B^T B)^{-1} B^T Yn = \begin{pmatrix} \hat{a} \\ \hat{u} \end{pmatrix}$$

doi:10.4304/jcp.5.5.703-708

$$Y_n = \begin{pmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(n) \end{pmatrix} B = \begin{pmatrix} -\frac{1}{2}\left[x^{(1)}(1)+x^{(1)}(2)\right] & 1 \\ -\frac{1}{2}\left[x^{(1)}(2)+x^{(1)}(3)\right] & 1 \\ \vdots & \vdots \\ -\frac{1}{2}\left[x^{(1)}(n-1)+x^{(1)}(n)\right] & 1 \end{pmatrix} \quad (4)$$

Sending $\hat{a},\hat{u}$ to the differential equations

$$x^{(1)}(k+1) = \left[x^{(0)}(1) - \frac{\hat{u}}{\hat{a}}\right]e^{-\hat{a}k} + \frac{\hat{u}}{\hat{a}}, (k = 0,1.2...) \quad (5)$$

Progression decrease the result to $x^{(0)}$

$$\begin{aligned} \hat{x}^{(0)}(k+1) &= \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k) \\ &= (1-e^{\hat{a}})(x^{(0)}(1) - \frac{\hat{u}}{\hat{a}})e^{-\hat{a}k}, (k = 0,1,2...) \end{aligned} \quad (6)$$

III. SVM ALGORITHM

SVM(Support vector machines) method is the effective practice to statistical learning theory, it is based on VC theory and Structural Risk Minimization, the optimum fitting between model complexity and learning ability based on finite sample information was established to reach the best generalization, it significantly better than the neural network method based on empirical risk minimization,it also can better solve the practical problems in the small sample, nonlinear and high dimensional. It has simple result, global optimization and better generalization ability in characteristics:

(1)It realizes the SVM principle, also minimizes the generalization error upper boundary, so it has the better upper boundary ability.

(2) Compared with neural network method, SVM has the less free parameters. There are three free parameters in SVM algorithm, but it should be selected by subjective.

(3)Neural network cannot always meet the global optimal solution, and it can easily fall into local optimal solution. In SVM algorithm, training SVM equivalent to solute a two convex planning problem with nonlinear constraint,so the solution is unique, global and optimal[1-3].

The disadvantage of SVM is that it can't determine which knowledge is redundancy. The core theory of SVM is making the input variables map to the high dimensional feature space H by nonlinear mapping $\varphi(\bullet)$, the optimum decision function is established by structure risk minimization principle, then the kernel function in original space is used to replace the dot product operation in high dimensional feature space.The optimum regression function in high dimension space H for SVM is as follows:

$$f(x) = w \bullet \varphi(x) + b \quad （7）$$

w——vector, $w \in R^k$

b——intercept, $b \in R$

At present, SVM includes ε-SVR and v-SVR. This paper used the v-SVR to explain. The v-SVR can be transformed into an optimization problem:

$$\frac{\min}{w,b,\xi_i^-,\xi_i^+}\frac{1}{2}w^T w + C\left[v\varepsilon + \frac{1}{l}\sum_{i=1}^{l}\left(\xi_i^- + \xi_i^+\right)\right]$$

$$s.t.\begin{cases} y^i - f\left(x^i\right) \le \varepsilon + \xi_i^- \\ f\left(x^i\right) - y^i \ge \varepsilon + \xi_i^+ \\ \xi_i^-, \xi_i^+ \ge 0 \quad i = 1,\cdots,l \end{cases} \quad （8）$$

C is penalty factor, v is the number of support vector machine, ε is insensitive loss function, $i = 1,\cdots,l$

The dual problem is as follows:

$$\frac{\min}{\alpha,\alpha^*}\frac{1}{2}\left(\alpha-\alpha^*\right)^T Q\left(\alpha-\alpha^*\right) + y^T\left(\alpha-\alpha^*\right)$$

$$s.t.\begin{cases} e^T\left(\alpha-\alpha*\right) = 0 \\ e^T\left(\alpha+\alpha^*\right) \le Cv \\ o \le \alpha_i, \alpha_i^* \le C/l \quad i = 1,\cdots,l \end{cases} \quad (9)$$

$Q_{ij} = K\left(x_i,x_j\right) \equiv \varphi\left(x_i\right)^T \varphi\left(x_j\right), K\left(x,y\right)$ is kernel function, $\varphi(\bullet)$ is mapping function in high dimension space.

The common kernel functions are as follows:

（1）linear kernel function

$$k(x,y) = (x,y)$$

（2）polynomial kernel function

$$K(x,y) = (x \bullet y + 1)^d, d = 1,2,\cdots$$

（3）RBF kernel function

$$K(x,y) = \exp\left(-\frac{\|x-y\|^2}{\sigma^2}\right)$$

（4）Laplace Function Kernel

$$K(x,y) = \prod_{i=1}^{n} e^{-u|x_i-y_i|}$$

（5）Radial Basis Function Kernel

$$K(x,y) = e^{-g\cdot\sum_{i=1}^{n}(x_i-y_i)^2}$$

（6）Cauchy Function Kernel

$$K(x,y)=\prod_{i=1}^{n}\frac{1}{1+u\left|x_i-y_i\right|^2}$$

$$u\in R^+$$

The approximation regression function is

$$f(x)=\sum_{i=1}^{l}\left(-\alpha_i+\alpha_i^*\right)K(x_i,x)+b \qquad (10)$$

The SVM structure is showed in Fig.1, $\alpha_i-\alpha_i^*$ is neural weight, $x_1,x_2,\cdots,x_m$ is input vector, y is output vector.

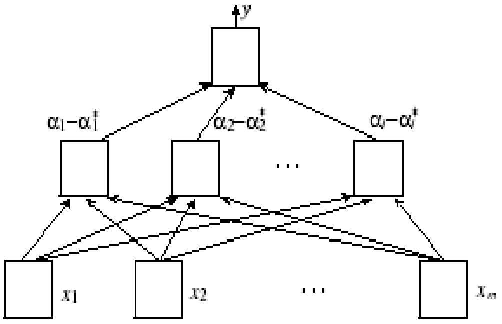


Figure1. the structure of SVM

IV. THE IMPROVED MARKOV MODEL

As the state probability that stochastic process on t_0 is known, the state probability more than t_0 only related to t_0, but it independent the state probability before t_0, we called this process is Markov process.

Assumed the state space of random sequence $\{x(n), n=1,2,...\}$ is E, if any integer $n_1,n_2,...,n_m\left(0\le n_1<n_2<...<n_m\right)$ and any natural number k meet

$$p\left\{X(n_m+k)\right\}=p\left\{X(n_m+k)=j\,\middle|\,X(n_m)=i_m\right\}$$

$i_1,i_2,..i_m,j\in E$, then random sequence $\{x(n), n=1,2,...\}$ is markov chain. It can be showed as follows:

$$x(n)=x(0)p^n \quad ,p=\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mn} \end{bmatrix}$$

$x(n)$ is the state probability vector in time n, $x(0)$ is the state probability vector in initial time, p is probability transfer matrix，the each row is a probability vector, it shows the probability that system state E_1 of i row transfer to other states[10-11].

Based on calculation, we can get the relative error (%) δ, according to δ, the four status E1, E2, E3, E4 are as follows:

Status	Value limit
E1	$\delta=(0,1]$
E2	$\delta=(-1,0]$
E3	$\delta=(1,6]$
E4	$\delta=(-6，-1]$

Suppose the total sample numbers is N, the transition probability from p to q is

$P_{pq}=\frac{N_{pq}}{N}$, N_{pq} is the transfer times from p to q

The probability status of system error can be determined by markov chain:

$\delta^*=\frac{\delta_{up}+\delta_{down}}{2}$, the power load forecasting value $F=F\bullet\left(1\pm\delta^*\right)$.

Due to the optional determination of divide upper and lower bound in an interval by the traditional Markov modification method, the problem of errors corrected without measure often occurs. So we proceed with selection method of upper and lower bound, round the error of one position before forecast position, and δ_{up} is supposed to be the absolute value of the rounded error, δ_{down} equals to 0.01, the sign before δ^* is determined by the state interval. With these works, the error modification can be more accurate, it is put into practice and turns out to be good effect.

V. DEMONSTRATION

In order to verify the forecast effect of the proposed method, the historical data of power consumption for a city is used to predict by Gary and SVM method, and then modify with posterior-error-test and Markov, the forecast results are gained, and a comparative verification was made at last. Details are showed as table 1.

TABLE I.
THE ELECTRICITY CONSUMPTION

Year	Electricity consumption	Year	Electricity consumption
	Billon kwh		Billon kwh
1995	40	2002	56.3
1996	40.8	2003	59.9
1997	49	2004	
1998	52	2005	
1999	56.9	2006	
2000	58.8	2007	
2001	56	2008	

A. Gray forecasting

The data from 1995 to 2003 are used as training samples, the data are normalized firstly, 5 former years data of power consumption are set to be input samples. The load data of 6 year which as output samples is used to rolling training by GM(1.1) method.

$$\hat{a}=-0.039131, \hat{u}=44.19$$

$$X(k+1)=(1-e^{\hat{a}})(x^0(1)-\frac{\hat{u}}{\hat{a}})e^{-\hat{a}k}$$

Obtained:
x(9+1)=63.82
x(10+1)=66.36
x(11+1)=69.01
x(12+1)=71.76
x(13+1)=74.63

Do Posterior-error-test to the above model, the process is:

mean residual:

$$\bar{\varepsilon}=\frac{1}{n}\sum_{k=1}^{n}\varepsilon(k)=\frac{1}{n}\sum_{k=1}^{n}\left[x^{(0)}(k)-\hat{x}^{(0)}(k)\right]$$

error variance:

$$S_1^2=\frac{1}{n}\sum_{k=1}^{n}\left[x^{(0)}(k)-\bar{x}\right]^2$$

Posterior-error-ratio C:

$$C=\frac{S_2}{S_1}$$

Infinitesimal error probability P:

$$P=P\{|\varepsilon(K)-\bar{\varepsilon}|$$

Then we utilize Posterior-error-ratio c, infinitesimal error probability P to evaluate the current model:

C=0.4621 evaluation result：good
p=0.8750 evaluation result：good
The predictive value of the next 5 times:
X(t+1)=63.81506
X(t+2)=66.36172
X(t+3)=69.01002
X(t+4)=71.76400
X(t+5)=74.62788
Qmin=-5.86244

TABLE II.
THE FORECAST ELECTRICITY CONSUMPTION CONTRAST

Serial number	True value	Predictive value	Fitting value	error
X(2)	40.80	46.66	-5.86	-14.37
X(3)	49.00	48.52	0.48	0.97
X(4)	52.00	50.46	1.54	2.96
X(5)	56.90	52.47	4.43	7.78
X(6)	58.80	54.57	4.23	7.20
X(7)	56.00	56.75	-0.75	-1.33
X(8)	56.30	59.01	-2.71	-4.82
X(9)	59.90	61.37	-1.47	-2.45

Continue to model residual series. The first analysis results of residual series:

Model parameters
$\hat{a}$=-0.155623 $\hat{u}$=2.083293
x(t+1)=19.249261exp(0.155623t)-13.386825

TABLEIII.
THE FORECAST ELECTRICITY CONSUMPTION CONTRAST TABLEII

Serial number	True value	Predictive value	Fitting value	error
X(2)	40.80	44.04	-3.24	-7.94
X(3)	49.00	52.31	-3.31	-6.76
X(4)	52.00	54.41	-2.41	-4.64
X(5)	56.90	55.63	1.27	2.23
X(6)	58.80	54.17	4.63	7.87
X(7)	56.00	53.13	2.87	5.12
X(8)	56.30	57.33	-1.03	-1.83
X(9)	59.90	63.79	-3.89	-6.49

The evaluation of the current model
C=0.4250 evaluation result：good
p=0.8750 evaluation result：good
The predictive value of the next 5 times:
X(t+1)=69.20921
X(t+2)=73.65131
X(t+3)=78.51421
X(t+4)=83.85570
X(t+5)=89.74280
Qmin=-8.96337
Continue to model residual series.
The 2nd analysis results of residual series:
Model parameters
$\hat{a}$=-0.117045 $\hat{u}$=1.645604
x(t+1)=23.022937exp(0.117045t)-14.059566

TABLEIV.
THE FORECAST ELECTRICITY CONSUMPTION CONTRAST TABLEIII

Serial number	True value	Predictive value	Fitting value	error
X(2)	40.80	37.94	2.86	7.02
X(3)	49.00	49.80	-0.80	-1.64
X(4)	52.00	55.61	-3.61	-6.95
X(5)	56.90	59.69	-2.79	-4.91
X(6)	58.80	57.47	1.33	2.27
X(7)	56.00	52.37	3.63	6.49
X(8)	56.30	54.34	1.96	3.49
X(9)	59.90	62.54	-2.64	-4.41

The evaluation of the current model

C=0.3745 evaluation result：good
p=1.0000 evaluation result：very good
The predictive value of the next 5 times:
X(t+1)=67.53763
X(t+2)=72.88516
X(t+3)=78.76591
X(t+4)=85.25164
X(t+5)=92.42505
Qmin=-4.34446
Continue to model residual series.
The 3rd analysis results of residual series:
Model parameters
$\hat{a}$ =0.039097 $\hat{u}$ =5.098622
x(t+1)=-126.066exp(-0.039097t)+130.41

TABLE V
THE FORECAST ELECTRICITY CONSUMPTION CONTRAST TABLEIV

Serial number	True value	Predictive value	Fitting value	error
X(2)	40.80	38.43	2.37	5.82
X(3)	49.00	47.24	1.76	3.58
X(4)	52.00	53.68	-1.68	-3.23
X(5)	56.90	61.20	-4.30	-7.55
X(6)	58.80	61.60	-2.80	-4.76
X(7)	56.00	55.01	0.99	1.77
X(8)	56.30	53.19	3.11	5.52
X(9)	59.90	59.28	0.62	1.03

The evaluation of the current model
C=0.3514 evaluation result: good
p=1.0000 evaluation result: very good
The predictive value of the next 5 times:
X(t+1)=66.72859
X(t+2)=71.94056
X(t+3)=77.69095
X(t+4)=84.05132
X(t+5)=91.10418
Qmin=-4.64647

B. SVM forecasting

Likewise，selecting the data from 1995 to 2003 are used as training samples, the data are normalized firstly, The load data of 6 year which as output samples is used to rolling training by SVM method, inner product function RBF is adopted to be kernel function, parameters are set as this: σ^2=30, C=100, ε=0.001. Through LIBSVM algorithm, we get load forecast values from 2004 to 2008. According to the former 5 years' data and their influencing factors, we select seven years' data from 2002 to 2008 in the table below. They are showed as table VI. .

TABLE VI.
ELECTRICITY CONSUMPTION PREDICTION RESULTS BY SVM

Year	Real data	SVM forecasting value	Relative error	status
2002	62.7	61.32	-0.022	E4
2003	65.914	62.59	-0.0504	E4
2004	59.13	60.13	0.01691	E1
2005	62.22	60.29	-0.031	E3
2006	74.812	72.91	-0.0254	E4
2007	86.686	82.21	-0.0516	E4
2008	88.64	89.43	0.00891	E3

Step one: Error analysis
We adopt relative error to analyze forecast errors:

$RE = \frac{y_i^* - y_i}{y_i}$, calculation results are showed as table VI.

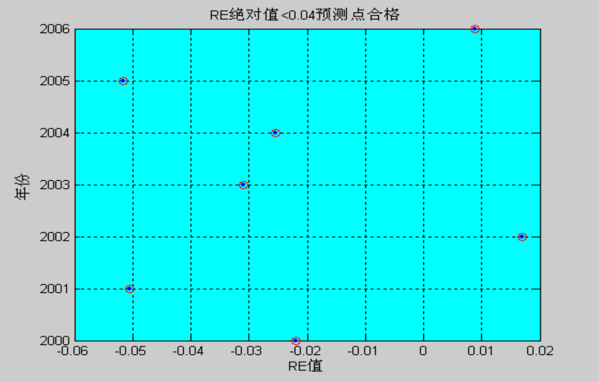


Figure2. Combination prediction relative errors

We determine qualified rate of forecasting mainly with the standard of 4% (it is a qualified forecasting position while $|RE|$ <4%), and determine forecast precision by average relative error simultaneously. The calculation results are obtained as this: average error of SVM is 0.0221; qualified rate is 71.43%, so they belong to a high level. We can see that the absolute value of RE to the forecast value are mainly between 0 and 0.04 from figure 2, this further illustrates the feasibility of the proposed method. For further optimization of forecast results, now introduce an improved Markov process to modify them.

Step two: The markov correcting

The transition probability matrix devised by relative error statement of SVM forecasting is

$$p = \begin{bmatrix} 0 & 0 & 0.17 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.17 \\ 0.17 & 0 & 0.17 & 0.32 \end{bmatrix}$$

Take the load data of 2006 for an example, according to the state probability vector of 2005 as $(1 \quad 0 \quad 0 \quad 0)^t$, using the formula we get the state probability vector of 2005, $\begin{pmatrix}0 & 0 & 0.17 & 0\end{pmatrix}^t$, it is obviously that the largest probability of load forecast error in state of E3 is 0.17, calculating the forecast value of 2005 in formula,

$$F=F\bullet\left(1+\delta^*\right)=60.29*\left(1+0.01\right)=60.9 \text{ Billon kwh}$$

The monitoring load data of 2005 is 62.22 billion kWh, the value corrected by Markov is 60.9 billion kWh, and the value forecasted by SVM method is 60.29 billion kWh. It is proved that the adoption of Markov error correction method can obviously increase forecast precision, and ensure the stability of forecasting.

Select the best value of gray prediction when p gradually reaches 1.0000 through posterior-error-test:

The result of forecast load which is 71.9 billion kWh in 2005 comparing with the true value 62.22 billion kWh，we can find that the result of the modified SVM is more approached to the true value comparing with the gray prediction of Posterior-error-test

VI. CONCLUSION

The method of Posterior-error-test is utilized to optimize gray model, and the optimization of C, P are obtained. Under these conditions, we can obtain the best predictive value.

If we have the same impact factors, we can obtain the fittest values which have a min error compare with the real results. The values are got by the SVM model which modified by Markov residual error.

Analyzing the reasons of comparatively large error on gray model, we can take the following improving measures:

（1）Further optimizing original series, such as exponential weighting, moving average;

（2）Select optimization data of C, such as median value, or final value;

（3）improving model: for example, we use other methods to define but not use the median of x(0)(k) and x(0)(k+1).

Sum up, we use the method of SVM to predict the long-medium power load forecast. But there are some uncertain factors of parameters, such as C, other variables. Therefore, the problem which how to improve the intelligent algorithm must be studied.

REFERENCES

[1] Niu Dongxiao, Liu Da. “Support Vector Machine Models Optimized by Genetic Algorithm for Hourly Load Rolling Forecasting”. *Transactions of China Electro technical Society*, 2007, (6), pp.148-153.

[2] Niu Dongxiao,Gu Zhihong.etc. “Study on Forecasting Approach to Short-term Load of SVM Based on Data Mining” [J]. *Proceedings of the CSEE*, 2006, (9), pp.6-12.

[3] Yang Jinfang,Zhai Yongjie. “Time series prediction based on support vector regression”.*Proceedings of the CSEE*, 2005,(9),pp.110-114.

[4] O. Habiballah, R. Ghosh-Roy and M. R. Irving. “Markov chains for multipartitioning large power system state estimation networks”.*Electric Power Systems Research*, 1998, (3),pp. 135-140.

[5] D. I. Jones and M. H. Lorenz. “An application of a Markov chain noise model to wind generator simulation”.*Mathematics and Computers in Simulation, 1986,(8),pp.391-402.*

[6] Dong Jizheng,Wang Huan.etc. “Application of Markov chain with weights to load forecasting”[J]. *Relay*, 2006,(3),pp.33-36.

[7] Niu, Dongxiao,Li, Jinchao. “Distribution center location decision-making based on PCA and support vector machine approach,”*Conference on Natural Computation*, 2007, pp.553-557.

[8] Niu, Dongxiao,Liu, Da,Chen, Guangjuan,Feng, Yi. “Support vector machine models optimized by genetic algorithm for hourly load rolling forecasting,” *Transactions of China Electrotechnical Society*, 2007, pp.148-153.

[9] Liu,Da,Niu,Dongxiao,Xing,Mian.“Day-ahead price forecast with genetic-algorithm-optimized support vector machines based on GARCH error calibration,”*Automation of Electric Power System 2007*, pp.31-34.

[10] Liu,Da,Niu,Dongxiao,Xing,Mian. “Day-ahead price forecast with genetic-algorithm-optimized support vector machines based on GARCH error calibration”. Dianli Xitong Zidonghua/Automation of Electric Power Systems, v 31, n 11, Jun 10, 2007, p 31-34+58

[11] NIU Dong-xiao etc. “Grey load forecasting models based on relational analysis of multi-factor”, Journal of North China Electric Power University，2006.(3):91－92.

Li Wei: (1968.12.14--), Gaobeidian, Hebei province, graduate from Heibei university at Baoding Hebei province in 1991, undergraduate from north china electric power university at Baoding Hebei province in 2006,D.candidate in north china electric power university at Beijing in 2007, major field of Technical and economic management, Research Area is power load forecast .

[1]Application of Improved Ant Colony Clustering in the Power Load Forecasting, journal Beijing institute of technology, 2007.12

[2]A New Intelligent Method for Power Load Forecasting, ISDA2006,2006

[3]Improved Genetic Algorithm–GM(1,1) for Power Load Forecasting Problem, DRPT2008,2008

Current and previous research interests is power forecast load.

Zhang Zhen-gang: (1977.03.25--), Xingtai, Hebei province, graduate from north china electric power university at Baoding, Hebei province in 2002, Postgraduate in north china electric power university at Baoding Hebei province in 2008, major field of Technical and economic management, Research Area is power load forecast .

Personalized Knowledge Acquisition through Interactive Data Analysis in E-learning System

Zhongying Zhao
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
Graduate School of Chinese Academy of Sciences, Beijing, China
Email: zy.zhao@sub.siat.ac.cn

Shengzhong Feng[1], Qingtian Zeng[2], Jianping Fan[1], and Xiaohong Zhang[1]
[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[2]College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao, China
Email: sz.feng@siat.ac.cn, qtzeng@sdust.edu.cn ,{jp.fan, xh.zhang}@siat.ac.cn

***Abstract*—Personalized knowledge acquisition is very important for promoting learning efficiency within E-learning system. To achieve this, two key problems involved are acquiring user's knowledge requirements and discovering the people that can meet the requirements. In this paper, we present two approaches to realize personalized knowledge acquisition. The first approach aims to mine what knowledge the student requires and to what degree. All the interactive logs, accumulated during question answering process, are taken into account to compute each student's knowledge requirement. The second approach is to construct and analyze user network based on the interactive data, which aims to find potential contributors list. Each student's potential contributors may satisfy his/her requirement timely and accurately. Then we design an experiment to implement the two approaches. In order to evaluate the performance of our approaches, we make an evaluation with the percentage of satisfying recommendations. The evaluation results show that our approaches can help each student acquire the knowledge that he/she requires efficiently.**

***Index Terms*—E-learning, knowledge acquisition, knowledge requirement, potential contributor**

I. Introduction

Personalized support becomes even more important, when e-Learning takes place in open and dynamic learning and information networks [1, 2, 3]. Personalized knowledge acquisition, is one of the most important phases for realizing user-adaptive or personalized e-learning systems. It involves several aspects. The first is to acquire what knowledge the student requires. The second problem is to find who can provide the related knowledge to satisfy student's requirement. Others include how to offer the knowledge, in what forms and what time. In this paper, we aim to solve the first two problems.

Interactive Question Answering (QA) system, which can be seen as virtual seminar, has been embedded into an e-learning system to improve learning performance [4, 5, 6]. During this system, students communicate their knowledge in the form of posing questions, selecting questions to answer and browsing others' questions and answers (Q&A). The e-learning system can store all these interactive logs into the data base in the form of question table, answer table and user table. All of these historical data contain a tremendous amount of information about the users' requirements and relations.

In this paper, we propose two approaches to achieve personalized knowledge acquisition. The first approach aims to mine what topics (what kind of knowledge) the user requires and to what degree. All his/her interactive logs, including posing question, answering questions and browsing answers, are taken into account to compute the knowledge requirement. This approach, however, does not imply whom a student can turn to when he/she has knowledge requirements. And the tightness of relations between students is also not reflected although it is very important for users to acquire knowledge. Thus, our second approach is to construct a user network for all the users in e-learning system. The user network describes each student's potential contributors list and the relation strength, which can improve the personalized knowledge acquisition.

Compared with others' excellent research results, the work in this paper is a supplement to achieve personalized E-learning, especially in personalized knowledge acquisition.

The remainder of the paper is organized as follows. In section 2, we discuss the related work. Section 3 describes the framework for personalized knowledge acquisition. Section 4 presents the approach for mining user's knowledge requirement. The construction and analysis of the user network is addressed in section 5, which aims to find the potential contributors. Section 6 combines our two approaches to achieve personalized knowledge acquisition. To evaluate our approaches, we design the experiment and evaluation in section 7. Section 8 concludes the whole paper and discusses the future work.

II. Related Work

doi:10.4304/jcp.5.5.709-716

To realize a user-adaptive or personalized e-learning system, user model and modeling are two of the key problems [7]. User model can be built based on user's behavior, the contents of a web page or both. A human behavior based user model can be learned by observing the user's actions such as web log file, path, clicking and downloads frequency [8-10]. A divisive hierarchical clustering (DHC) algorithm to group terms is proposed by Kim H.R.& Chan P.K.[11]. Dwi H. Widyantoro et al. propose a three-descriptor model to represent a user's interest categories and an adaptive algorithm to learn the dynamics of the user's interests through positive and negative relevance feedback [12].Trajkova and Gauch build user profiles automatically from the web pages and focuses on improving the accuracy of the user profile based on concepts from a predefined ontology [13]. Liu lu and Wu lihua model user's interest and value related characteristics in recommender systems [14]. A vision-based approach to detect user's interests on web pages is proposed in literature [15]. Considering that users' interests may be inferred from what they read and how they interact with documents, Rajiv Badi et al present models for detecting user interest based on reading activity [16].

With the development and application of Web technology, interactive question answering system has also become an active research area. An interactive question answering system named BuyAns is developed for personalized E-learning implementation [4]. HITIQA is designed to allow intelligence analysts and other users of information systems to pose questions in natural language and obtain relevant answers or the assistance [17]. With fully-implemented interactive QA system named FERRET, Sanda Harabagiu et al achieved a surprising performance by integrating predictive questions into the context of a QA dialogue [18]. Chun-Chia Wang et al proposes a repository-based Question Answering system for collaborative E-learning [19].

Although many Web-based learning techniques have been proposed to assist adaptive/personalized Web-based learning, few researches have attempted to acquire user's knowledge requirement and potential contributors. Personalized knowledge acquisition, including the acquisition of knowledge requirement and potential knowledge contributors, plays very important role in promoting personalized learning efficiency. We have done a lot of work in user modeling, including interest mining [20, 21], knowledge requirement acquisition [22]. This paper is an expanded version of the conference paper [21]. Some modifications of our QA system are described in this paper. Another approach for capturing the potential knowledge contributors is proposed, which can accelerate the personalized knowledge acquisition combining with our previous works.

III. FRAMEWORK

As illustrated in Fig.1 is the framework to describe the whole process for personalized knowledge acquisition within e-learning systems. This framework can be embedded into any e-learning system as an assistant component to realize user-adaptive or personalized learning or instruction. The main components contained in the framework include the following:

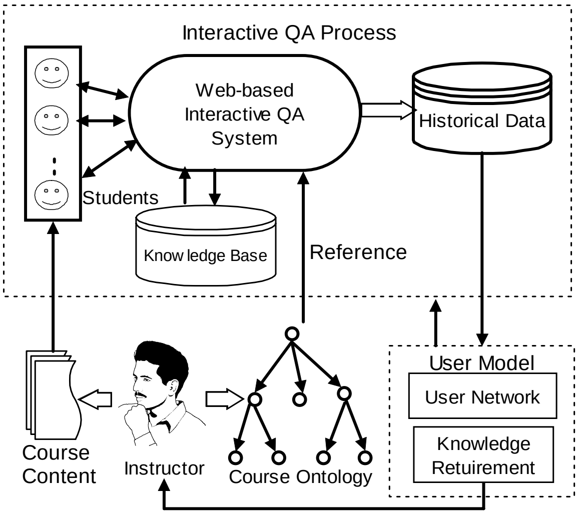


Figure 1. The framework for personalized knowledge acquisition within e-learning systems

(1) Course ontology (or concept hierarchy): The course ontology is predefined by the instructor with suitable granularity and scale. It presents the structure of the course content and provides a reference for QA board structure. An example for Artificial Intelligence (AI) course ontology is described in Fig.2 and [20]. In this course ontology, there are 8 leaf terms (T1-T8), which corresponds to the QA boards.

(2) Interactive QA process: Based on the predefined course ontology, the corresponding board structure of the user-interactive QA system can be generated. Within their favorite boards, users can post their most urgent questions, browse their favorite answers, and select others' questions to answer. All these interactive QA data can be recorded and accumulated to historical data base to acquire student's knowledge requirement and potential contributors. For the questions posed by students, some of them are answered by knowledge base. And others require students or instructor to give answers, which will be discussed in section IV.

(3) User modeling: It is a critical part for personalized knowledge acquisition. This process includes acquiring each student's knowledge requirement and constructing user network. According to the historical data, we can compute each student's knowledge requirement in every question. Then the mapping relation between question and topic is used to compute his/her knowledge requirement in each topic. The historical interactive data are also used to construct user network. The user network describes the relations between students during the QA process, which can be used to find the potential contributors for each student.

User modeling is done once a week. Then each student's knowledge requirement and user network can be used to guide the next interactive QA process and achieve personalized knowledge acquisition. Furthermore, based on student's knowledge requirement model, the instructor can adjust her/his teaching materials and offer personalized helps to each student.

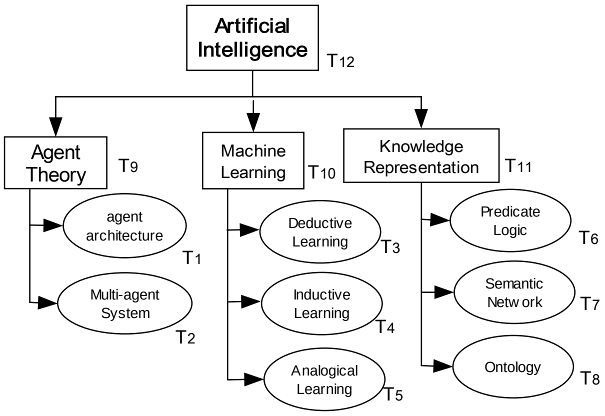


Figure 2. The concept hierarchy for AI course ontology

IV. Mining user's knowledge requirement

During the interactive QA process, students can propose questions, browse others' answers and select their favorite questions to answer. While browsing the answers given by others, the student is asked to submit his/her rate through clicking the button "satisfying" or "not satisfying". The button "satisfying" corresponds to 1 while the button "not satisfying" corresponds to 0.

For the question posed by a student, he/she has to wait for others to answer. In order to shorten the time spent in waiting for answers, we classify the questions into $PlainQ$ and $Probe\text{-}intoQ$. $PlainQ$ is answered by knowledge base, while $Probe\text{-}intoQ$ is answered by other students or their instructors.

A. Classification of questions

(1) $PlainQ$ refers to the question that has an exact or accurate answer, such as querying a fact or a definition. Based on the questions proposed by students, we extracted the corresponding pattern. The pattern set for $PlainQ$ is given as follows.

PSplain = { *What is* (*the definition of*) [*concept*]?
Which | *What* [*areas* | *fields*] *does* [*a concept*] *cover* | *include* | *contain*?
When does the [*concept*] *be proposed* | *presented* | *given* (*in*) (*some fields*)?
Who proposes | *presents* | *gives the* (*concept of*) [*concept*]?...},
where
[] *represents compulsory field*, () *represents optional field*.

If a question satisfies one pattern, it will be submitted to knowledge base. And the answer will be given by knowledge base immediately. With the enlarging of knowledge base, new patterns are being added to the pattern set in order to satisfy user's requirement.

(2) $Probe\text{-}intoQ$ refers to the question that can be analyzed and answered from many different perspectives. E.g. What conclusions can be drawn ……? How to understand……? What are the reasons for……?

For this kind of question, it is often dealt with as the following. (1) It is answered by some helpful students. For this circumstance, the questioner has to spent long time waiting for answers. (2) The instructor is charge of every question. Since the instructor is the expert for this course or domain, his/her answers are often accurate and make questioner satisfied. A disadvantage is that, however, with students and questions increasing, the instructor will become the bottleneck. (3) All the others are encouraged to answer the proposed questions. But only some/few of them can satisfy the questioner's knowledge requirement. That is, a lot of junk information/answers are made during this process. To improve the interactive QA efficiency and achieve personalized knowledge acquisition, we should find the potential contributors for each student. This will be discussed in section V.

B. Historical data

In the interactive QA system, students interactively exchange their knowledge by questioning, answering, and browsing. All these historical data can be stored in the form of user table, question table, and answer table. The follows are the attributes of each table.

TABLE 1. User table

UID	*OperType*	*QID/AnsID*	*Judge*

TABLE 2. Question table

QID	*QTopic*	*WhoPro*	*QType*	*NumAns*

TABLE 3. Answer table

AnsID	*QID*	*WhoAns*	*NBrow*	*NSatisfy*	*NNoSatisfy*

For the user table, UID is the primary key. $OperType$ represents the operation the student makes, and $OperType \in \{\text{proposing, browsing, answering}\}$. $QID / AnsID$ refers to the question or answer given by this student. The selection of QID or $AnsID$ is determined by $OperType$. $Judge$ means whether the student is satisfied with the given answer. $Judge = NULL,\ if\ \ OperType = \text{answering}$. Otherwise, $Judge = \begin{cases} 1, \text{if the students is satisfied} \\ 0, \text{if the students is not satisfied} \end{cases}$.

For the question table, QID is the primary key. $QTopic$ refers to the board that the question belongs to, and $Qtopic \in \{T_i \mid i = 1, 2, ..., 8\}$, T_i is the leaf term of course ontology. $WhoPro$ is the UID that proposes this question. $QType$ denotes the type of the question, and $QType \in \{PlainQ, Probe\text{-}intoQ\}$. $NumAns$ denotes the number of the answers given to this question.

For the answer table, $AnsID$ is the primary key. QID denotes the question that the answer belongs to. $WhoAns$ represents the UID that gives this answer. $NBrow$ means the number that the answer is browsed. $NSatisfy$ is the total number of $Judge = 1$ for this answer. Similar to the above, $NNoSatisfy$ is the total number of $Judge = 0$ for this answer. And according to the description of interactive process in this section, we can derive that $NBrow = NSatisfy + NNoSatisfy$.

C. Acquiring user's knowledge requirement

From the historical data, we can get each student's interactive logs through operations on the table. In order to compute each student's knowledge requirement from the interactive logs, we firstly define $QASet_i$ of student i.

$$QASet_i = \left\{(Q_j, AnS_j, p_{i,j}, f(B)_{i,j}, f(A)_{i,j}) \mid j = 1, 2, ..., n\right\},$$

where

(1) Q_j is a question.

(2) AnS_j is the set of answers to Q_j.

$$|AnS_j| = \begin{cases} 0, \ if\ Q_j \text{ is a question without answer.} \\ 1, \ if\ Q_j \in PlainQS. \\ m\ (m > 1), if\ Q_j \text{is a } Probe\text{-}intoQ. \end{cases}$$

(3) $p_{i,j} = \begin{cases} 0, if\ Q_j \text{ is proposed by student } i. \\ 1, if\ Q_j \text{ is not proposed by student } i. \end{cases}$

(4) $f(B)_{i,j}$ denotes the number of answers browsed by student i to Q_j.

(5) $f(A)_{i,j}$ denotes the number of answers given by student i to Q_j.

By common sense, the student who posts question Q_j usually requires the corresponding knowledge more urgently than those only browsing. For question Q_j, the knowledge requirement of student i is defined as follows:

$$KR_i(Q_j) = \frac{1}{\pi}\arctan(\alpha p_{ij} + \beta(f(B)_{ij} - Meanf_{B_i}) + \gamma f(A)_{i,j}) + \frac{1}{\pi}\arctan(\beta Meanf_{B_i}) \qquad (1)$$

where

(1) α, β, γ are parameters and $\alpha > \beta > \gamma$,

(2) $Mean\ f_{B_i}$ is the average number of browsing of student i

Since the structure board of the QA system is generated according to the course ontology and all the questions and answers are distributed in each board. From the question table (Tab.2), we can clearly know the topic that each question belongs to. Then we can compute each student's knowledge requirement about each leaf term of course ontology, i.e. student's knowledge requirement about each sections of course content.

$$KR_i(T_k) = \frac{1}{m}\sum_{j=1}^{m} KR_i(Q_j), \text{ where} \\ (Q_j, T_k) \in MR,\ m = \left|\left\{Q_j \mid (Q_j, T_k) \in MR\right\}\right| \qquad (2)$$

V. CAPTURING POTENTIAL CONTRIBUTORS

A. User Network Construction

a. All questions without users involved.

From the data base described in section Ⅳ, we can construct the mapping relations between all the questions and answers (Fig.3).

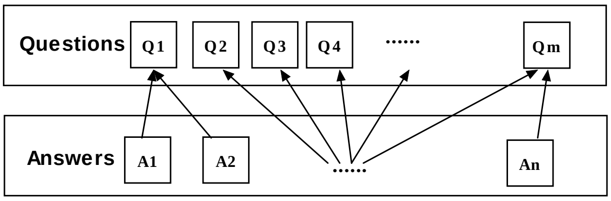


Figure 3. Mapping relations between questions and answers

b. One question with users involved.

In order to construct user network, we firstly consider the situation for one question Q_j. From the database, we can select all the users related to this question, including those who propose/answer this question and those who browse the answers to this question. To intuitively represent the relations between users, questions, answers, we have defined the following rules:

(1) Circles are employed to denote users while squares are used to denote questions or answers;

(2) There is a direct edge from U_i to Q_j if user U_i proposes the question Q_j;

(3) There is a direct edge from U_i to A_j if U_i gives the answer A_j;

(4) There is a direct edge from A_j to U_i if U_i browses the answer A_j and considers it good;

According to the database and the rules defined above, we can construct a graph for the question Q_j. Two examples are shown in Fig.4 and Fig.5. Fig.4 represents two special situations. One is that no one has given answer to Q_j proposed by U_p. The other is that the Q_j posed by U_p is a *PlainQ* and it has been answered by data base. In Fig.5, U_p proposes Q_j and then $U_1, U_2,$ and U_3 give 4 answers in total. U_2 and U_4 browse U_1 's answer and judge it good. U_6 browses the answers given by U_2 and U_3 and consider them valuable. U_5 also browses U_3 's answer and gives a good estimate.

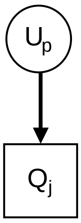


Figure 4. The question Q_j without answer

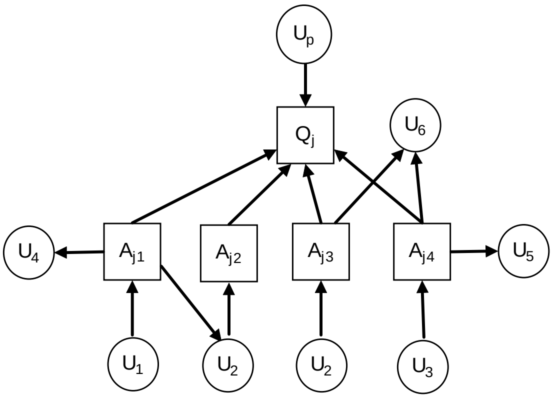


Figure 5. The question Q_j with many answers

To clearly describe the relations between users involved in one question, it is necessary for us to delete the question(s) and answers (Q&A). We present the following steps for Q&A elimination.

U_i walks to the Q_j along the direction of the edge and replace the Q_j if there is a direct edge from U_i to Q_j ;

(2) U_i walks to A_j along the direction of the edge and replace the A_j if there is a direct edge from U_i to A_j ;

(3) If there are N(N>1) direct edges from U_i to U_j , all the edges should be merged and marked a weight N.

(4) For other edges without weight, we assign 1 as the weight.

Following the 4 steps, we can eliminate all the questions and answers. Fig.4 is transformed to a isolated point named U_p . Fig.5 is converted to a weighted graph as illustrated in Fig.6. That is a simple user network since only one question is taken into account.

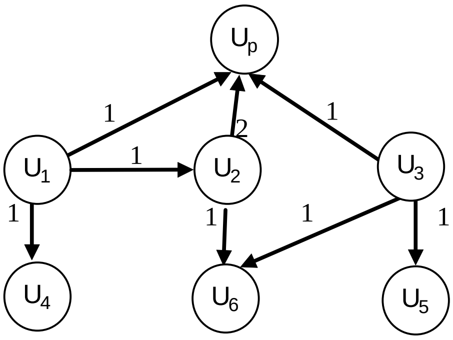


Figure 6. The user network for one question Q_j

To facilitate the computer processing, we adopt matrix to denote the graph above. The user network for Q_j is defined as $UM_{N*N}(T_k)^j$, where Q_j belongs to the topic T_k .

$$UM_{N*N}(T_k)^j = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1N} \\ w_{21} & w_{21} & & w_{2N} \\ \dots & \dots & \dots & \dots \\ w_{N1} & w_{N2} & \dots & w_{NN} \end{bmatrix}, \text{ where}$$

(1) N is the total number of users in the interactive QA system.

(2) w_{ij} is the weight marked on the directed edge from U_i to U_j .

c. One topic with users involved

The user network for all questions can be considered to be a merge problem, since we have described the construction of simple user network for one question above. For the topic T_k , we can easily construct the matrix, denoted by $UM_{N*N}(T_k)$.

$$UM_{N*N}(T_k) = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1N} \\ w_{21} & w_{21} & & w_{2N} \\ \dots & \dots & \dots & \dots \\ w_{N1} & w_{N2} & \dots & w_{NN} \end{bmatrix}$$

$UM_{N*N}(T_k)$ can be computed from the following formula:

$$UM_{N*N}(T_k) = \sum_{j=1}^{j=M} UM_{N*N}(T_k)^j , \tag{3}$$

where M is the number of questions on board T_k in QA system.

The following algorithm presents the whole process for constructing the user network.

Algorithm. Constructing user network for the interactive QA system

```
INPUT: historical data base (user table, question table,
       answer table)
OUTPUT: user network for each topic $UM_{N*N}(T_k)$
Step 1: define and initialize a N*N matrix $UM_{N*N}(T_k)$,
        $UM_{ij}(T_k) = 0$ (i,j=1,2, … ,N);
Step 2: for every question in topic $T_k$, do
        If $U_i$ give an answer to $U_j$, $UM_{ij}(T_k) += 1$;
        If $U_j$ browses $U_i$ 's answer and judge it good,
        $UM_{ij}(T_k) += 1$;
Step 3: output $UM_{N*N}(T_k)$
```

B. User network analysis

According to the construction and definition of user network, we can see that the $UM_{N*N}(T_k)$ is essentially a directed graph. As to $UM_{N*N}(T_k)$, w_{ij} refers to the weigh marks on the directed edge from U_i to U_j . It represents how many contributions U_i makes to U_j . And,

(1) If $w_{ij} \ge 1$, we say U_i is a potential contributor to U_j , and U_j is a potential beneficiary.

(2) $\sum_{j=1}^{j=N} w_{ij}$ represents how many contributions U_i makes to others.

(3) $\sum_{i=1}^{i=N} w_{ij}$ means the total gains that U_j obtains from others.

In (1) above, we consider U_i a potential contributor because it has made contributions to U_j in the past. It may help U_j in the future but not always. For the same reason, U_j is a potential beneficiary. Therefore, the user network $UM_{N*N}(T_k)$ describes and quantifies the strength of relations between users under the topic T_k. Here, the relation refers to contributor – beneficiary.

From the *j*th column of $UM_{N*N}(T_k)$, we can get each U_j 's potential contributors list by sorting U_i according to the w_{ij} decreasing.

VI. Personalized Knowledge Acquisition

From the first approach, we can obtain each student's knowledge requirement, including which topics they require, to what degree, who has the same or the similar knowledge requirement with each other. The second approach aims to find each student's potential contributors, which can help students acquire the knowledge that they require quickly and accurately. In this section, we combine the two approaches to achieve personalized knowledge acquisition.

In our interactive QA system, each student registered is assigned an agent as the assistant. The agent plays very important role in the process of personalized knowledge acquisition. It is in charge of the call of our two approaches once a week. And it maintains the user's knowledge requirement computed from the first approach. The potential contributors list is also maintained by the agent after sorting based on the decrease of w_{ij} (i=1,2, … ,N). Each agent can browse the board and recommend others' new questions and answers to its own user.

To achieve personalized knowledge acquisition, all the results from the first and second approach are taken into account. The higher student's knowledge requirement degree, the more we select from the potential contributors list. In this paper, we adopt the 5-3-2 rule. Take the topic T_j for example, U_i 's agent selects K contributors from U_i 's potential contributors list. And,

$$K = \begin{cases} 5, \text{ if } \delta_1 \le KR_i(T_j) \le 1 \\ 3, \text{ if } \delta_2 \le KR_i(T_j) < \delta_1 \\ 2, \text{ if } 0 < KR_i(T_j) < \delta_2 \end{cases}$$

That means, if $\delta_1 \le KR_i(T_j) \le 1$, the agent selects 5 contributors with the highest weight from U_i 's potential contributors list. When U_i proposes a question on T_j, its agent will recommend it to the 5 users and invite them to give answers. Once a new answer appears, the agent notifies its user to browse. Meanwhile, the agent forward the corresponding link to U_i 's potential beneficiaries in order to satisfy their implicit knowledge requirement. Then users browse the questions and answers and submit their evaluations. Al l these behaviors are recorded in the database, which will be used to compute or update each user's knowledge requirement and user network.

VII. Experiment and Evaluation

A. Experiment design and result

To evaluate the performance of our approaches, we conduct an experiment with AI course ontology. As to AI, we only select 3 chapters and 8 sections to simplify the experiment. The concept hierarchy is shown in Fig.2.

10 students majoring in the Artificial Intelligence are invited to attend this experiment. During the first four weeks of the experiment, we encourage every student to answer others' questions actively and helpfully. And all the students are encouraged to post their urgent questions on their favorite boards. During the interactive process, each student pays more attention to his/her favorite questions. Once the answer that a student browses satisfies his/her requirement, he/she clicks the “satisfy” button to submit the judge. Otherwise, “not satisfy” button is clicked.

After four weeks, we have collected and stored abundant Q/A historical data in the form of user table, question table and answer table. Then each student's agent calls our two approaches. The first proposed approach is applied to compute the knowledge requirement of each student shown in Fig.7. The second approach is invoked to construct user network for each topic. An example of user network for T_8 is shown as follows:

$$UM_{N*N}(T_8) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 7 & 0 & 5 & 1 & 0 & 8 & 0 & 0 \\ 0 & 4 & 0 & 0 & 1 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 6 & 1 & 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 5 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

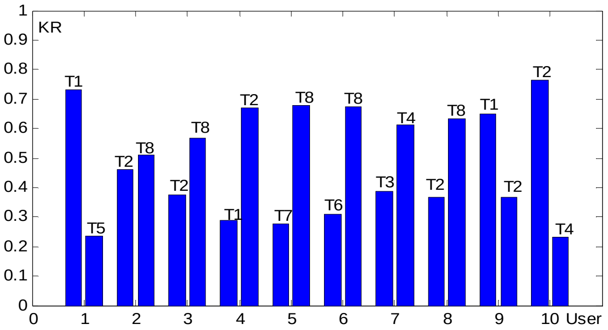


Figure 7. The knowledge requirement of each student

B. Evaluation

In this subsection, we make an evaluation to examine the performance of our two approaches. The performance is measured in terms of percentage of satisfying recommendations (Perc of SatRec). A satisfying recommendation is defined as a positive recommendation such that, after receiving and browsing it, student is satisfied with the recommendation and click “satisfy” to submit his/her evaluation. Thus, the percentage of satisfying recommendations is the total number of recommendations divided by the number of satisfying recommendations.

We set $\delta_1 = 0.6$, $\delta_2 = 0.3$ in this evaluation and the potential contributors are selected based on the rule described in section Ⅵ. All the results computed from the first and second approach are taken into account to achieve personalized knowledge acquisition. According to the student's knowledge requirement $KR_i(T_j)$ computed from the first approach, the agent selects K contributors on top of the potential contributors list. When U_i proposes a question on T_j, its agent will recommend it to the K students and invite them to give answers. Once a new answer appears, the agent notifies its user to browse. Meanwhile, the agent forward the corresponding link to U_i 's potential beneficiaries in order to satisfy their implicit knowledge requirement. Then users browse the questions and answers and submit their evaluations.

If our approaches can achieve personalized knowledge acquisition, that means expressing student's knowledge requirement and helping him/her acquire knowledge precisely and quickly, the percentage of satisfying recommendations should be keep higher. The percentage of satisfying recommendations of the 10 students is calculated every day. After 60 days for the content recommendation, the percentages of satisfying recommendations are shown in Fig.8. We can see that all the percentages are stable and always higher than 0.7. Further more, the changing spectrum is narrowing down with time passing by. Thus, the experimental results show that the two proposed approaches realize personalized knowledge acquisition to some degree.

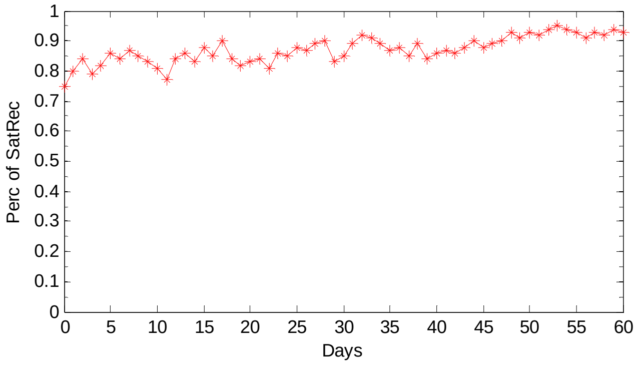


Figure 8. The Students' percentage of satisfying recommendations within 60 days

VIII. CONCLUSION

Personalized knowledge acquisition is very important for learning efficiency. In this paper, we propose two approaches to achieve personalized knowledge acquisition in interactive QA system. The first approach is to acquire the knowledge requirement of the users from their historical QA logs. It aims to mine what kind of knowledge or what topic is required by students. And it also quantifies the urgency of each student's knowledge requirement. The second approach presents the construction of user network based on students' interactive history. It aims to find the potential contributors and improve the personalized knowledge acquisition quickly and accurately.

An experiment is conducted to implement our proposed approaches, obtaining the knowledge requirement and potential contributors list of each student. The evaluation combined with the experiment results indicate that our approaches realize personalized knowledge acquisition in interactive QA system to some degree.

In our future work, we will extend or modify our interactive QA system, so that users can upload their resources for knowledge sharing. Then the potential beneficiary can learn from his/her potential contributors and acquire more knowledge that he/she implicitly requires.

ACKNOWLEDGMENT

This research is partly supported by National High-tech Research and Development Program of China (863 Program, No.2006AA01A114), National Science Foundation of China (No.60603090, 90718011 and 50875158), the Research Foundation of Shandong Educational Committee of China (No.J08LJ77), and the Taishan Scholar Program of Shandong Province. The authors would like to thank all the students for their efforts in our experiment.

REFERENCES

[1] Chen, C., Lee, H., and Chen, Y. Personalized e-learning system using item response theory. Computers & Education. 2005(44), pp.237-255.

[2] Dolog, P., Henze, N., Nejdl, W., and Sintek, M. Personalization in distributed e-learning environments. In Proceedings of the 13th international World Wide Web Conference on Alternate Track, 2004, pp.170-179.

[3] Thyagharajan, K. & N., R. Adaptive content creation for personalized e-learning using web services. Journal of Applied Sciences Research, 2007, 3(9), pp.828-836.

[4] Dawei Hu, Wei Chen and Qingtian Zeng et al. Using a user-interactive QA system for personalized E-learning. International Journal of Distance Education Technologies, 2008(6), pp.1-22.

[5] J. C. Hung, C.-S. Wang, C.-Y. Yang, M.-S. Chiu, and G. Yee. Applying word sense disambiguation to question answering system for e-learning. In Proceedings of the 19th international conference on advanced information networking and applications, 2005(1), pp.157-162.

[6] Wang, Y., Wang, W., and Huang, C. Enhanced semantic question answering system for e-learning environment. In Proceedings of the 21st international conference on Advanced Information Networking and Applications Workshops. 2007(2), pp.1023-1028.

[7] Brusilovsky, P., Sosnovsky, S., and Shcherbinina, O. User Modeling in a Distributed E-Learning Architecture. User Modeling, 2005(3538), pp.387-391.

[8] Kazunari Sugiyama, Kenji Hatano, Masatoshi Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In Proceedings of the 13th international conference on World Wide Web, 2004, pp.675-684.

[9] Feng Qiu, Junghoo Cho. Automatic identification of user interest for personalized search. In Proceedings of the 15th international conference on World Wide Web. 2006, pp.727-736.

[10] Himanshu Sharma, Bernard J. Jansen. Automated evaluation of search engine performance via implicit user feedback. The 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2005, pp.649-650.

[11] Hyoung R. Kim and Philip K. Chan. Learning implicit user interest hierarchy for context in personalization. In Proceedings of Intelligent User Interfaces. 2003, pp.101-108.

[12] D.H. Widyantoro, T.R. Ioerger, and J. Yen. Learning user interest dynamics with a three-descriptor representation. Journal of the American Society of Information Science and Technology, 2001, 52(3), pp.212-225.

[13] J. Trajkova and S. Gauch. Improving ontology-based user profiles. In Proceedings of the Recherched' Information Assiste par Ordinateur (RIAO), 2004, pp.380-389.

[14] Liu Lu, Wu lihua. User modeling for personalized recommender system. Tsinghua science and technology. 2005(10), pp.772-777.

[15] Yin Liu, Wenyin Liu and Changjun Jiang. User interest detection on web pages for building personalized information agent. Advances in web-age information management. 2004 (3129), pp. 280-290.

[16] Rajiv Badi, Soonil Bae, J. Michael Moore et al. Recognizing user interest and document value from reading and organizing activities in document triage. In Proceedings of the 11th international conference on intelligent user interfaces. 2006, pp.218-225.

[17] S. Small, T. Liu, N. Shimizu, T. Strzalkowski. HITIQA: an interactive question answering system a preliminary report. In Proceedings of the ACL workshop on multilingual summarization and question answering. 2003(12), pp.46-53.

[18] Sanda Harabagiu, Andrew Hickl, John Lehmann and Dan Moldovan. Experiments with interactive question-answering. In Proceedings of the 43rd annual meeting on Association for Computational Linguistics. 2005, pp.205-214.

[19] Wang, C. C., Huang, J. C., Shih, T. K., & Lin, H. W. (2006). A repository-based question answering system for collaborative e-learning. Journal of Computers. 2006, 17(3), pp. 55-68.

[20] Yongquan Liang, Zhongying Zhao, Qingtian Zeng. Mining User's Interest from Reading Behavior in E-learning System. The 8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD), 2007(2), pp. 417-422.

[21] Zhongying Zhao, Shengzhong Feng, Yongquan Liang, Qingtian Zeng, Jianping Fan. Mining user's interest from interactive behaviors in QA System, First International Workshop on Education Technology and Computer Science. 2009(12), pp.1025-1029.

[22] Qingtian Zeng, Zhongying Zhao, Yongquan Liang. Course Ontology-based User's Knowledge Requirement Acquisition from Behaviors within E-Learning Systems, Computers & Education, In Press, Corrected Proof, Available online 28 May 2009, doi:10.1016/j.compedu.2009.04.019.

Zhongying Zhao is born in Shandong Province, 1983. She is a PhD candidate both in Shenzhen Institute of Advanced Technology and Institute of Computing Technology, Chinese Academy of Sciences, China. Her research interests include data mining, parallel and distributed computing, high performance computing, and personalized e-learning.

Shengzhong Feng is a professor at the Shenzhen Institute of Advanced Technology, CAS. His research focuses on parallel algorithms, grid computing and bioinformatics. Specially, now his interests are in developing novel and effective methods for digital city modeling and application. Before came to SIAT, CAS, he had been working in the Institute of Computing Technology, CAS, and participated in the Dawning supercomputer research and development. He graduated from the University of Science and Technology of China in 1991, and received his PhD from Beijing Institute of Technology in 1997.

Qingtian Zeng is a professor at the College of Information Science and Technology, Shandong University of Science and Technology. His research interests are Workflow Mining, Domain Ontology and Petri net. He has published papers in leading international journals such as IEEE MASC. He received the support from Natural Science Foundation of China and the Excellent Young Scientist Foundation of Shandong Province of China. He is serving on the editorial board of the Journal of Software Engineering and Asian Journal of Information Management, and was the Program Co-Chair of KSEM07, GCC07, WaGe2007, and KSEM09.

Jianping Fan is a professor at the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. His research interests include high performance computing, Grid computing, and computer architecture. He took part in and developed Dawning I, Dawning 1000, Dawning 3000, Dawning 4000 and other series of Dawning supercomputers. He has published more than 70 papers and 1 book. He also acquired 11 pending or issued patents. Professor Fan has received many awards, such as outstanding award of CAS science and technology progress, first price of national science and technology progress， first price of Beijing science and technology progress, and the outstanding young scientist of CAS. He has actively participated in various professional activities. He served as editor of journal of computer research and development, Advisor of 11th five years science and technology development plan of ministry of information industry, and general chair of HPC China2007 and GCC2008.

Xiaohong Zhang is a PhD candidate in Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China. Her research interests include massive data management, data access pattern analysis and data mining.

Application of Chaos Mind Evolutionary Algorithm in Antenna Arrays Synthesis

Jianxia Liu
Taiyuan University of Technology, Taiyuan, China
Email: tyljx@163.com

Nan Li and Keming Xie
Taiyuan University of Technology, Taiyuan, China
Email: linanwd@163.com

Abstract—**Mind evolutionary algorithm (MEA) uses 'similartaxis' operation and 'dissimilation' operation to imitate the human mind evolution to processes optimization, overcoming the prematurity and improving searching efficiency. But it has several defects: the generation of the initial population is blind, random and redundant; the addition of naturally washed out temporary subpopulations is monotonous; existing searching modes easily to fall into local convergence. This paper proposed Chaos Mind Evolutionary Algorithm. Two chaotic sequences produced in different ways bring adequate diversity to the population. As a result, the searching area is widened. Chaotic Mind Evolutionary Algorithm is used in antenna array synthesis in this paper. Computer simulations show that Chaos Mind Evolutionary Algorithm can be applied in optimization problems of uniformly-spaced linear array and the optimization result is better than that obtained from Genetic Algorithm.**

Index Terms—**mind evolutionary algorithm, chaotic optimization, similartaxis operation, dissimilation operation, pattern synthesis**

I. INTRODUCTION

With the rapid development of the telecommunications industry, the electromagnetic environment of space is increasingly deteriorating, electromagnetic interference is enhancing and the quality of communication is declining. To solve these problems, smart antenna which has the ability of low side-lobes, strong directional and anti-interference has been a great deal of concern. The antenna pattern synthesis problem becomes a hot research. Antenna array synthesis means in a given antenna radiation pattern or antenna performance, design antenna array element number, element spacing, elements' current amplitude and phase distribution. For some antenna arrays with given element number and element spacing, this problem is to find every elements' excitation current amplitude and phase distribution.

Because the objective function or constrains of antenna optimization problems are multi-parameter, nonlinear, non-differentiable and even discontinuous, so the traditional numerical optimization methods which based on gradient optimization technology can not effectively achieve the satisfactory results of the project. Intelligent algorithm has become a powerful tool for optimal design because of its strong global search and the search is not dependent on the specific problems' gradient information and searching space's information. In recent years, Intelligent Algorithm gets access to a wide range of applications and the development in microwave technology and antenna design.

Genetic algorithm (GA) is a kind of intelligent algorithm which often applied to study synthesis of antenna arrays in recent years. GA was applied in synthesis of antenna arrays in 1994 first by J. M. Johnson and Y. R. Samii[1]. But GA is easy to be trapped in part optimum value, and convergence speed reduces obviously in later searching stage.

Mind evolutionary algorithm (MEA) is brought forward based on thinking of human mind development [2]. MEA simulated the similartaxis and dissimilation phenomenon in human society and resolved the problem of prematurity and low convergence speed of traditional Intelligent Algorithm to a certain extent. In this paper, chaos is incorporated into MEA to construct a Chaotic MEA (CMEA), where the parallel population-based evolutionary searching ability of MEA and chaotic searching behavior are reasonably combined. The algorithm not only has good searching guide but also make the best of chaos's ergodicity, so that the algorithm has higher convergence rate and better searching ability. CMEA is applied in optimization problems of uniformly-spaced linear array. Simulation results and comparisons demonstrate the effectiveness and efficiency of CMEA in antenna synthesis.

II. MIND EVOLUTIONARY ALGORITHM

Many real optimization problems can be formulated as the following functional optimization problem.

$$\min f(x)$$
$$a_i \le x_i \le b_i, i = 1,2,3\cdots, n \qquad x = (x_{1,}x_2,\cdots x_n) \quad (1)$$

This research is supported by the Nature Science Foundation of Shanxi Province of China.（No.2008011027-4 and No.20051037）, corresponding author: Nan Li, linanwd@163.com.

doi:10.4304/jcp.5.5.717-724

where f is the objective function, a_i and b_i are lower and upper bounds for the variable x_i , and n is the dimensions of the variable vector x .

The aggregation of all individuals in every generation in MEA is called a population; a population is divided into some subpopulations. There are two kinds in subpopulations: superior subpopulations and temporary subpopulations. Superior subpopulations record the winners' information in global competition; temporary subpopulations record the middle process in global competition. The billboard provides the chance for the communication of the individuals and the subpopulations. There are three basic kinds of information in billboard: the sequence number, action and score of the individual or the subpopulation. The score is the valuation that the environment evaluates to the action of the individual or subpopulation. The individuals in subpopulations paste their information on local billboard. And the global billboard is used to paste the subpopulations' information.

MEA has two important operations 'similartaxis' operation and 'dissimilation'. In all subpopulations, the process that the individuals compete for the winners is called similartaxis. In the whole solution space, the process of each subpopulation competing for the global winner and ceaselessly prospecting for new point in the solution space is called dissimilation. Similartaxis exploits the part information that system gets from the environment, quickly search for the local optimum. However, the dissimilation operation searches in the whole solution space and choose better individuals as centers to create new temporary subpopulations. If a subpopulation can't produce new point in the similartaxis process, the subpopulation has been mature.

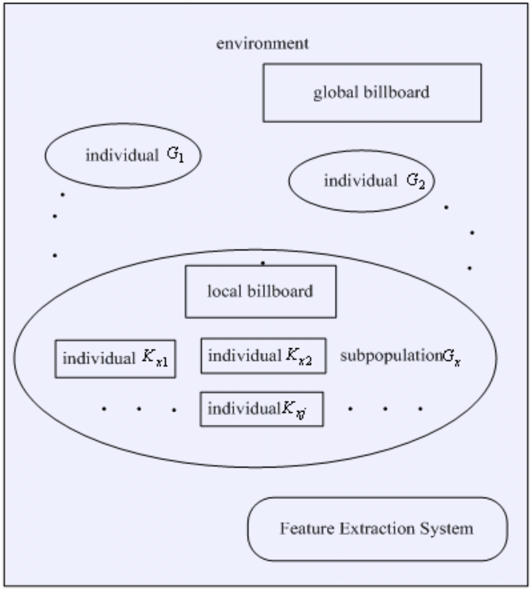


Figure 1. Framework of MEA

The simple MEA is described as following:

Step1 Set evolutionary parameters: population size, subpopulation size and conditions for end.

Step2 Initialization: scatter individuals composing initial population in whole solution space.

Step3 Similartax: individuals are produced by normal distribution with variance around each winner and the individual with highest score is the new winner replacing the old one in following steps.

Step4 Dissimilation: realize global optimization, some with lower score are washed out and replaced by new ones scattered at random in solution space.

Step5 Conditions for end: if the end conditions are filled, turn to step6; else repeat step3 and step 4.

Step6 Output evolutionary result, algorithm ends.

III. Chaos Optimization Algorithm

The phenomenon of chaos is the common phenomenon in the nonlinear dynamic systems. The chaos's behaviors are complex and similar to the random process, but have the inherent property of regularity. The chaos optimization algorithm is sensitively to the initial value, easily to jump out the local minimum, and quickly to search out the global optimization. The computation precision of chaos optimization algorithm is high. It has the property of global asymptotical convergence [4-6].

A. The Chaostis Characteristic of Logistic Mapping

Logistic mapping is the most typical model of Chaos Dynamics. It can be expressed as follow:

$$x_{k+1} = \mu \cdot x_k \cdot (1 - x_k)\,,\ n = 0,1,...,N \quad x_0 \in (0,1) \qquad (2)$$

where μ is the control parameter. Regard the finite difference eq. (2) as a dynamic system and it exhibits chaotic dynamics when $\mu = 4$ and $x_0 \notin \{0,0.25,0.5,0.75,1\}$. That is, when the control parameter $\mu = 4$, the system which doesn't have the stable solution at the completely chaotic state, and the chaos variable x_n ergodic in the scope (0,1). It also exhibits the sensitive dependence on initial conditions, which is the basic characteristic of chaos. A minute difference in the initial value of the chaotic variable would result in a considerable difference in its long time behavior. The track of chaotic variable can travel ergodically over the whole search space.

The probability density function of logistic mapping is

$$\rho(x) = \begin{cases} \dfrac{1}{\pi\sqrt{x(1-x)}}, 0 < x < 1 \\ 0, x \le 0, x \ge 1 \end{cases} \qquad (3)$$

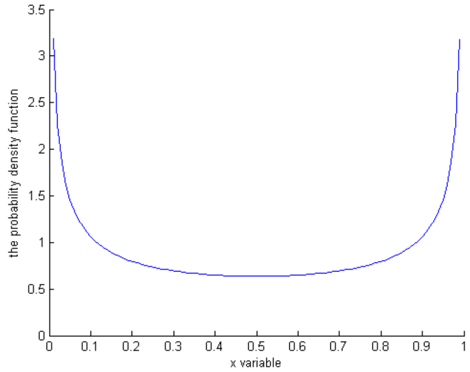


Figure 2. Logistic's probability density function

The curve shows that the probability is higher in [0,0.05] and [0.95,1]. As a result, the distribution of Logistic is non-uniform. If the points produced by chaotic mapping is the range of [0.05, 0.95], the searching time will be longer.

B. The Chaostis Characteristic of Tent Mapping

Tent map has uniform probability density, power spectral density and ideal related characteristics. Its formulate is

$$x_{k+1} = \begin{cases} 2x_k, 0 \le x_k \le 0.5 \\ 2(1-x_k), 0.5 < x_k < 1 \end{cases} \tag{4}$$

Its probability density function is

$$\rho(x) = 1 \tag{5}$$

Tent mapping has simple structure and good ergodic uniformity, more suitable for a large number of data processing sequences. It iterates faster than Logistic mapping. But there are small iterative cycle and unstable periodic point in tent mapping. It will make the iteration to the fixed point 0. In order to avoid iterates to fixed point, this paper uses the following method to improve tent mapping:

$$x(k+1) = \begin{cases} x(k)/0.4 & x(k) \le 0.4 \\ (1-x(k))/0.6 & x(k) > 0.4 \end{cases} \tag{6}$$

We can see that from Fig.3 the points generated by improved tent mapping are scattered uniformly in [0.1]. It overcomes its own lacks, such as small cycle and instability cycle points.

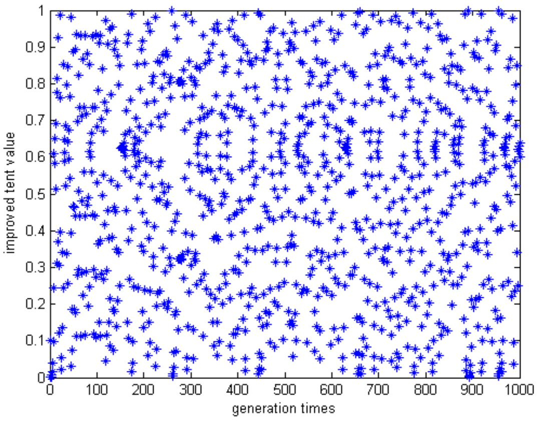


Figure 3. The chaos state of improved tent mapping

IV. CHAOTIC MIND EVOLUTIONARY ALGORITHM

Based on the proposed SMEA and the chaotic local search(CLS), a two-phased iterative strategy named Chaotic MEA (CMEA) is proposed, in which SMEA is applied to perform global exploration and CLS is employed to perform locally oriented search (exploitation) for the solutions resulted by MEA.

The method of improving the algorithm is: use improved tent chaotic sequence to generate initial population. And use logistic chaotic sequence to add washed out temporary subpopulations.

The randomness, ergodicity and the initial data's sensitivity of chaotic sequence ensure that the values will be uniformly distributed in the solution space. So it may be able to overcome data redundancy of the random sequence. On the other side, it increases the diversity of the population and expands the searching scope of the algorithm by using different chaotic sequence to add washed out temporary subpopulations.

To assess the performance of CEMA, this paper selects three typical functions that commonly used to test optimization algorithm to experiment, which has multi-peaks, non-raised and so on. The objective function is

$$F_1 = 100({x_1}^2 - x_2)^2 + (1-x_1)^2$$
$$-2.048 \le x_1, x_2 \le 2.048$$
$$F_2 = \left(4 - 2.1x_1^2 + x_1^4/3\right)x_1^2 + x_1x_2 + \left(-4 + 4x_2^2\right)x_2^2$$
$$-10 \le x_1, x_2 \le 10$$
$$F_3 = \frac{\sin^2\sqrt{\left(x_1^2 + x_2^2\right)} - 0.5}{\left[1.0 + 0.001\left(x_1^2 + x_2^2\right)\right]^2} - 0.5$$
$$-10 \le x_1, x_2 \le 10$$

F_1 is Rosenbrock function which has a global minimum. The global minimum position is (1, 1) and the value is 0. It is used to test the premature convergence of the algorithm. F_2 is Camel function which has six local minimums. The two global minimums are -1.031628, and their positions are (-0.0898, 0.7126) and (0.0898, -0.7126). F_3 is Schaffer function which has infinite local maximum . The only one global is maximum 1 and its position is (0, 0). The parameters of the experiment are: MEA and CMEA have the same large initial populations, the population size is 30, the subpopulation size is 18, the temporary subpopulation is 12, and termination time is 100.

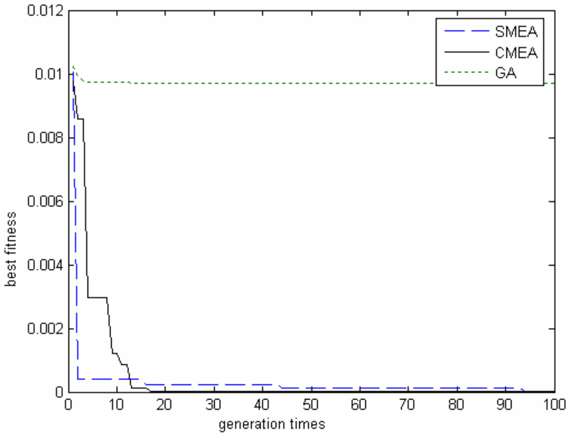


Figure 4. The convergence curve of Rosenbrock

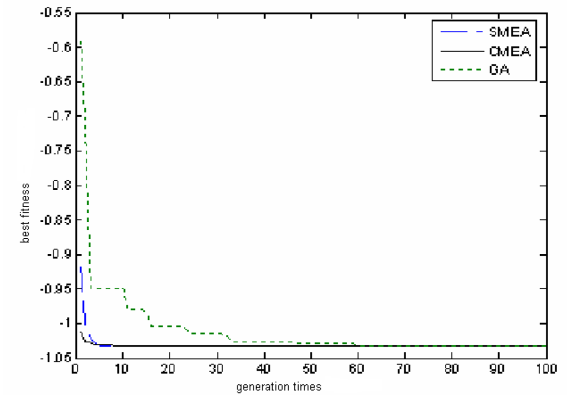


Figure 5. The convergence curve of Camel

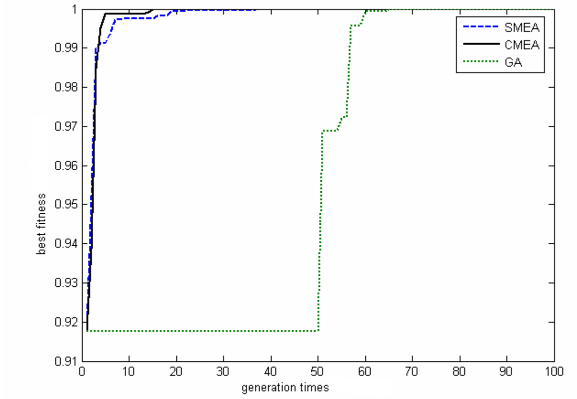


Figure 6. The convergence curve of Schaffer

TABLE I.
COMPARISON OF THREE FUNCTIONS' OPTIMIZATION RESULTS

function	algorithm	X_1	X_2	solution
Rosenbrock	MEA	0.9947	0.9990	1.8231e-6
	CMEA	0.9996	0.9992	8.0936e-7
	GA	0.9091	0.8309	3.0325e-3
Camel	MEA	0.0899	-0.7122	-1.031625035
	CMEA	-0.0898	0.7126	-1.031628033
	GA	-0.0876	0.7106	-1.031600471
Shaffer	MEA	0.0066	-0.0030	0.9975
	CMEA	0.0014	-0.0001	0.9981
	GA	0.0098	0.0098	0.9975

From Fig.4-6 and Table 1, it can be seen in the three kinds of functions optimization, MEA and CMEA are effective. They are both able to find the optimal solution. From the solutions' accuracy, the optimal values' accuracy gained by CMEA is higher than those gained by MEA and the solutions' positions are more accurately. This is because in different stages of the optimization process adopting different chaotic sequences to generate subpopulations. It's not only to ensure the uniform distribution of the subpopulations and also improves the quality of the individuals. Thereby enhance the convergence of the algorithm and the accuracy of the optimal value.

V. THE THEORY OF ANTENNA ARRAY SYNTHESIS

A. Traditional Methods of Antenna Array Synthesis

The antenna array synthesis is that given the radiation pattern of antenna array or given the antenna array's performance parameters, designing the elements' number, the space between elements, the amplitudes and phases of all the elements. For an antenna array with given elements' number and the space between elements, it is to optimize the amplitudes and phases of the array elements.

Generally speaking, antenna pattern synthesis can be classified into three categories. One group requires that the patterns exhibit a desired distribution in the entire visible region. This is referred to as beam-forming, and it can be accomplished using the Fourier transform and the Woodward-Lawson method. Another category requires that the antenna patterns possess nulls in desired directions. The method introduced by Schelkunoff can be used to accomplish this. A third group includes techniques that produce patterns with narrow beams and low side lobes.

There have been many classic methods in antenna array synthesis, such as Woodward method, Chebyshev polynomial method, Taylor synthesis method, Schelkunoff polynomial method and so on. Woodward method is that for a required radiation pattern, through sampling in different discrete location to achieve the expected pattern. But if the number of antenna array element is too large, the antenna radiation pattern gained by Woodward method is more ups and down, as it shows in Fig.7. Chebyshev polynomial method is that if the side-lobes' level is given, we can gain the narrowest main-lobe; if the width of the main-lobe is given, we can gain the lowest side-lobes' level. However it restricts all of the side-lobe on the same level, as it shows in Fig.8. It is not good for the whole antenna design. And if the element space is smaller than $\lambda/4$, Chebyshev polynomial method is not suitable for this kind of problem.

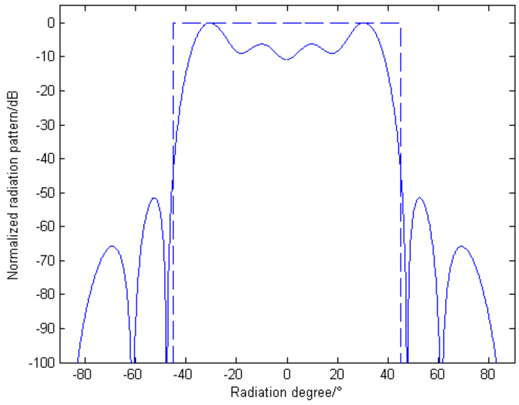


Figure 7. The antenna pattern gained by Woodwrd

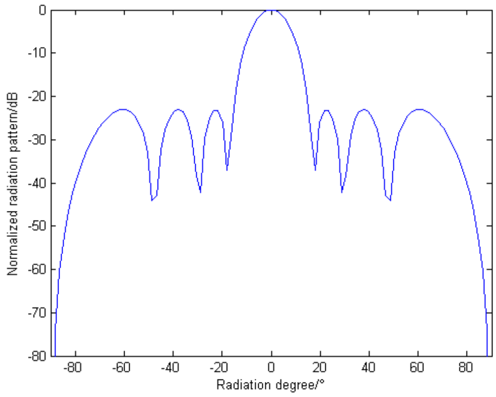


Figure 8. The antenna pattern gained by Chebyshev

This shows that classic methods of antenna array synthesis often have following characteristics:

• They are generally based on gradient search optimization.

• Most of these methods are pointed at special problems and their application scopes are small.

• Two problems can not be avoided. Firstly it must choose good initial data to ensure the achievement of the optimization objective. Finally in the solution space, there are special requirements for the continuity and differentiability of the objective function, such as requiring the objective function is continuous and differentiable in the solution space. However, the object functions for synthesis of array antennas usually have the characteristics of multi-parameters, non-differentiable even discontinuities.

Therefore classic gradient-based optimization methods are difficult to achieve satisfactory results.

B. The Mathematical Model of Antenna Array Synthesis

For a linear antenna array (shown as Figure9) with given elements and spacing arranged in axis as it shows in Fig.9, its pattern formula is

$$F(\theta)=\sum_{n=1}^{N} I_n \exp\left[j(n-1)kd\cos\theta+\varphi_n\right]$$

where I_n is the n element's amplitude; φ_n is the phase difference between adjacent elements; θ is the angle between the array axis and the ray; d is the space between the elements; $k=2\pi/\lambda$ is wave number.

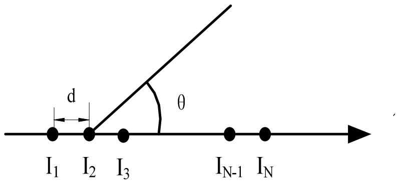


Figure 9. The model of uniform linear antenna

Let the antenna array's main-lobe point at θ_0 , then $\varphi_n=-(n-1)kd\cos\theta_0$, eq. (7) can be written as following

$$F(\theta)=\sum_{n=1}^{N} I_n \exp\left[j(n-1)kd(\cos\theta-\cos\theta_0)\right] \quad (8)$$

Let the pattern's imaginary part be zero, equation (8) can be written as following

$$F(\theta)=\sum_{n=1}^{N} I_n \cos\left[(n-1)kd(\cos\theta-\cos\theta_0)\right] \quad (9)$$

If N is even and the current's amplitudes is symmetrical, then the equation is

$$F(\theta)=\sum_{n=1}^{N/2} 2I_n \cos\left[\frac{2n-1}{2}kd(\cos\theta-\cos\theta_0)\right] \quad (10)$$

In antenna array, the symmetric array is usually used to cut down the parameters' number, then reducing the computing amount.

C. The General Objective Function of Antenna Array Synthesis

Antenna array synthesis is a multi-objective, multi-parameters and non-linear optimization problem. In engineering application, it is impossible to set null in every interference location to restrict interference. We can only choose some strong interference to generate null. At the same time, adopt design of low side-lobes level to restrict other interference from other direction. Considering the main-lobe position, the main-lobe width, side-lobes' level, null position and null depth, we can choose a general objective function [7]:

$$\begin{aligned} fitness &= w_1\cdot\frac{|\theta_0-\theta_{des}|}{180^0} \\ &+w_2\sum_{i=1}^{N} a_i\cdot|SLL_{\max}-SLL_{des}| \\ &+w_3\cdot\frac{\theta_{BWFN}-\theta_{BWFN\_des}}{180^0} \\ &+w_4\sum_{i=1}^{M} b_i\cdot\left|NULL_{\theta_I}-NULL_{\theta_i\_des}\right| \end{aligned} \quad (11)$$

where θ_0 is the main-lobe's position in experiment and θ_{des} is the required main-lobe's position; $SLL_{\max}$ is the highest side-lobes' level in experiment and SLL_{des} is the required side-lobes' level; θ_{BWFN} is the main-lobe's width in experiment and θ_{BWFN_des} is the required main-lobe's width; $NULL_{\theta_I}$ is the null depth at θ_i in experiment and $NULL_{\theta_I\_des}$ is the required null depth at θ_i. w_i is every objective's weight coefficient. Weight coefficients are very important for the whole design. They are directly related to the objectives' convergence trend and convergence rate. So we must analyze every objective's value and choose appropriate weight coefficients to balance the optimization rate of every optimization objective and obtain a best global optimal solution. Through analysis and computer simulation, we obtain the weight factors' general range: $w_1\in[0.3,0.5]$, $w_2\in[0.9,1.4]$, $w_3\in[0.5,0.8]$, $w_4\in[0.1,0.3]$.

VI. Simulation Results

Different methods, GA and CMEA, were investigated and compared with simulation solutions in order to assess the effectiveness and the flexibility of the proposed method. The experiment parameters of GA are: p_c=0.6, p_m=0.1. The experiment parameters of CMEA are: the subpopulation size is 18, the temporary subpopulation is 12. In two algorithms, the evolutionary generation is 100 and the population size is 30.

Example 1

For a antenna array with $d = \lambda/2$, $N = 12$, we request main lobe point at 0° , the shaped range is [-45°,45°], the normalized antenna pattern is $F(\theta) = 1$ in the shaped range. The simulation results are as Fig.10:

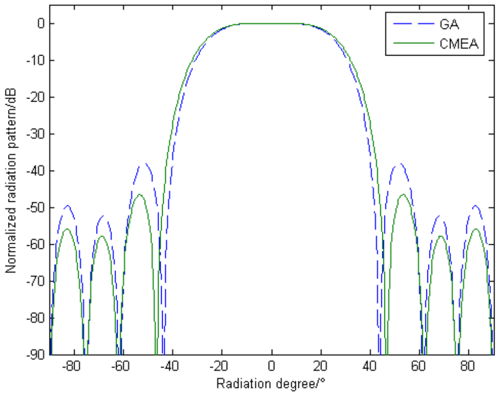


Figure 10. Radiation pattern of shaped beam

TABLE II.
COMPARISON OF OPTIMIZED RESULTS OF NORMALIZED AMPLITUDES

Number	GA	MEA
1	0.0098	0.0021
2	0.0325	0.0046
3	0.6636	0.4575
4	0.8344	0.6863
5	1.0000	0.9150
6	0.9955	1.0000
7	0.9955	1.0000
8	1.0000	0.9150
9	0.8344	0.6863
10	0.6636	0.4575
11	0.0325	0.0046
12	0.0098	0.0021

From Fig.10 and Fig.2, we can see that the antenna pattern gained by Woodward-Lawson, fluctuating about 10 dB in shaped range. But the optimization patterns gained by GA and CMEA are similar and satisfy the requirement. It is interesting to observe that the CMEA can make the better main-lobe, the less null level, as well as the side-lobe peak value is lower than GA.

Example 2

For a antenna array with $d = \lambda/2$, $N = 9$, we request the antenna pattern is $F(\theta) = \csc^2\theta$ in shaped range $[9°,30°]$, the side-lobes must be lower as possible as they can. Simulation results are as Fig.11:

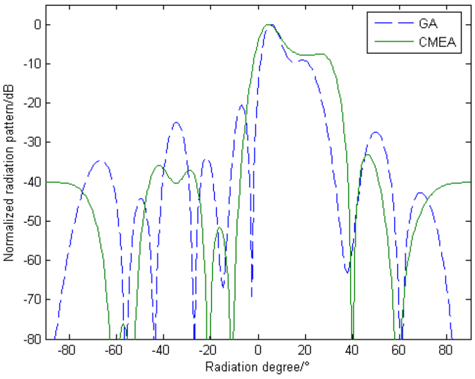


Figure 11. Radiation pattern of shaped beam

TABLE III.
COMPARISON OF OPTIMIZED RESULTS OF NORMALIZED AMPLITUDES AND PAHSES DIFFERENCE

N	GA		CMEA	
	amplitude	phase	amplitude	phase
1	0.9798	-47°	0.7869	-18°
2	1.0000	-55°	1.0000	-36°
3	0.9073	-87°	0.7183	-83°
4	0.6855	-84°	0.1439	-72°
5	0.2460	-75°	0.4034	-3°
6	0.0968	-67°	0.2303	-77°
7	0.6331	-90°	0.0001	-80°
8	0.1532	-47°	0.0001	-54°
9	0.1452	-70°	0.0883	-68°

From the antenna pattern, it can be seen that the pattern optimized by CMEA is good agreement with the desired radiation patterns while its side-lobes reduce greatly. The comparative side-lobe is 15dB lower than GA's. From their convergence curves, we can know that CMEA has faster convergence rate and better fitness function than GA. From Table 3 we can see that the optimized amplitudes and phases are very different because the optimization problem of array antenna is a multi-value problem which has the similar pattern with different amplitudes and phases of elements.

Example 3

For a antenna array with $d = \lambda/4$, $N = 16$, we request the main-lobe point at $\theta_0 = 0^0$, the width of main-lobe is 20°and the highest side-lobe level is -30dB. The simulation results are as Fig.12:

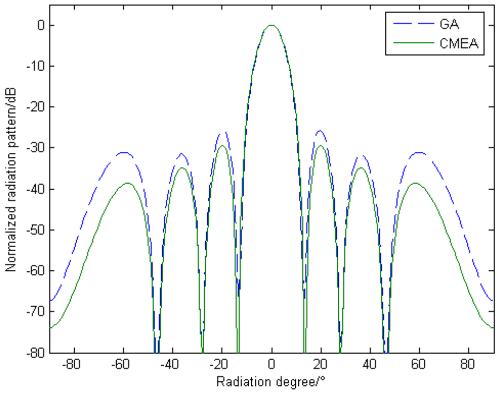


Figure 12. Radiation pattern of least side-lobe in designed range

TABLE IV.
AMPLITUDES OF ARRAY ELEMENTS

N	GA	CMEA
1	0.7497	0.4572
2	0.0000	0.7879
3	1.0000	0.7349
4	0.5484	0.1837
5	0.0971	0.7795
6	0.9537	0.6898
7	0.5596	0.3837
8	0.8846	1.0000
9	0.8846	1.0000
10	0.5596	0.3837
11	0.9537	0.6898
12	0.0971	0.7795
13	0.5484	0.1837
14	1.0000	0.7349
15	0.0000	0.7879
16	0.7497	0.4572

As it can be seen from Fig.12, the two kinds of algorithm can also satisfy the requirement that the main-lobe points at 0°, the width of main-lobe is 20°. However, the highest side-lobe's level of GA is -26dB, and in the results of CMEA every side-lobe's level is lower than that at the same direction in GA. The minimum can be achieved -40dB.

Example 4

For a antenna array with $d=\lambda/2$, $N=12$, we request the main-lobe point at $\theta_0 = 0^0$, the width of main-lobe is 10°, the highest side-lobe level is -20dB. And at 10, 20, 30, 35, 40, 60 form nulls which below -100dB. The simulation results are as Fig.13:

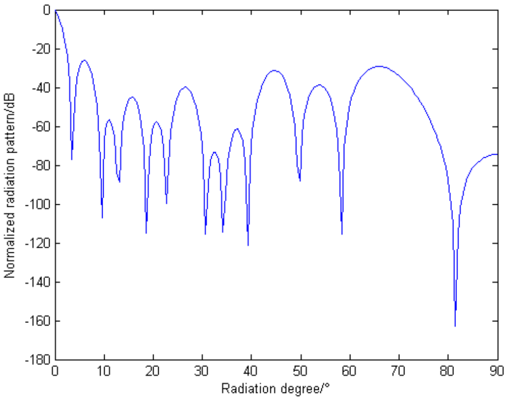


Figure 13. Radiation pattern of nulls in designed positions

TABLE V.
PHASES OF ARRAY ELEMENTS

N	1	2	3	4	5	6
phase	15°	49°	24°	12°	19°	40°
N	7	8	9	10	11	12
phase	25°	8°	54°	6°	59°	46°

In this optimization, we only choose phase as the optimization parameter. This is because the digital phase shifter technology has matured, and in phased antenna array it does not need to pay additional costs. So that in recent years this method is much more attractive. According to the experimental requirements, six locations need to achieve -100dB null depth, at the same time consider the main-lobe's position and the largest side-lobe level. The problem belongs to multi-objective optimization problem. For multi-objective optimization problems, it often not only has one global optimal solution. In this optimization, we adopt weighted method (choose different weights for different objectives) to design the objective. As Fig.13 shows, the antenna pattern gained by CMEA satisfies the multi-objective requirement.

VII. CONCLUSION

Point at the lacks of Mind Evolutionary Algorithm such as the generation of the initial population is blind; the addition of naturally washed out temporary subpopulations is monotonous, we integrated of Mind Evolutionary Algorithm and Chaos Optimization Algorithm with their respective advantages and proposed a new hybrid optimization algorithm. It uses improved tent chaotic sequence to generate initial population. And use logistic chaotic sequence to add washed out temporary subpopulations, increasing the diversity of the population and expanding the searching scope of the algorithm. CMEA is adopted to optimize the amplitude and phase of equal interval linear array. There is good agreement between the desired and calculated radiation patterns. The optimizing result is better than that of GA.

ACKNOWLEDGMENT

This research is supported by the Nature Science Foundation of Shanxi Province of China. （No.2008011027-4 and No.20051037）.

REFERENCES

[1] J. M. Johnson and Y. R. Samii, “Genetic Algorithm Optimization and its Application to Antenna Design”, *Antennas and Propagation Society International Symposium*, vol. 47, pp. 326-329, Jun 1994.

[2] C. Y. Sun, K. M. Xie. and M. Q. Cheng, “Mind Evolution Based Machine Learning Framework and New Development”, *Journal of Taiyuan University of Technology*, vol. 30, pp. 453-457, Sep 1999.

[3] G. Xie, J. L. Gao and K. M. Xie, “Fuzzy modeling Based on Rough Sets and Mind Evolutionary Algorithm”, *Journal of Electronic Measurement and Instrument*, vol. 1, pp. 109-112, Aug 2003.

[4] K. Keiji, “Stability extended delayed-feedback control for discrete time chaotic systems”. *IEEE Trans. On Circuits and Systems*, vol. 46, pp.1285-1288, Oct 1999.

[5] L. Chen, G. R. Chen, “Fuzzy modeling, prediction and control of uncertain chaotic systems based on time series”. *IEEE Trans. Circuits and Systems-I: Fundamental Theory and Application*, vol. 47, pp.1527-1531, Oct 2000.

[6] R. L. Devaney, *An Introduction to Chaotic Dynamical Systems*. Addison-Wesley, 2nd ed, 1989, New York.

[7] Y. Fan, R. H. Jin and B. Liu, “Study on the Objective Function for the Array Pattern Synthesis Based on Genetic Algorithm”. *Journal of Electronics and Information Technology*, vol. 27, pp. 801-804, May 2005.

[8] M. G. Lin, F. J. Liu and X. Q. Jiang, “A New Chaos Immune Optimization Algorithm”, *Journal of Hengyang Normal University*, vol. 28(3), pp.86-89, Jun 2007.

[9] J. J. Yang, J. Z. Zhou and J. Yu, “Particle Swarm Optimization Algorithm Based on Chaos Searching”, *Computer Engineering and Application*, vol. 16, pp. 69-71, Mar 2005.

[10] X. Zhou, Y. H. Hu and X. Q. Chen, “Chaos Genetic Algorithm and its Application in Function Optimization”, *Computer and Digit Engineering*, vol. 33, pp. 68-70, Dec 2004.

[11] W. L. Stutzman and G. A. Thiele, *Antenna Theory and Design (Second Edition)*. Beijing.Post & Telecom Press. 2006, pp.41-366.

[12] K.K. Yan and Y. Lu, “Sidelobe reduction in array-pattern synthesis using genetic algorithm”, *IEEE Trans. Antennas Propagat*., vol. 45, pp. 1117-1122, Jul 1997.

[13] J. X. Liu, F. Wang and K. M. Xie, “Mind Evaluation Algorithm Based on Chaos Searching”, *Computer Engineering and Applications*, vol. 44, pp. 37-39, Sep 2008.

[14] L. N. Yang, J. Ding and C. J. Guo, “Pattern Synthesis of Antenna Array Using Genetic Algorithm”, *Journal of Microwaves*, vol. 21, pp. 38-41, Apr 2005.

[15] R.L. Haupt, “Thinned arrays using genetic algorithms”, *IEEE Trans. Antennas Propagat*., vol. 42, pp. 993-999, Jul 1994.

[16] Y. X. Qiu and K.M Xie, “A New Method to Optimizing Fuzzy Control Rules”, *Journal of Taiyuan University of Technology*, vol. 35, pp. 254-256, May 2004.

[17] C. L. Wang and K. M. Xie, “Convergence of a New Evolutionary Computing Algorithm in Continuous State Space”, *Computer Math*., vol. 79, pp. 27-37, 2002.

[18] K. M. Xie, Y. G. Du and C. Y. Sun, “Application of the Mind-Evolution-Based Machine Learning in Mixture-Ratio Calculation of Raw Materials Cement”, *Proceedings of the 3rd World Congress on Intelligent Control and Automation*, pp. 132-134, 2000.

Jianxia Liu was born in China in July 1970. She got her Master of Engineering in electronic circuit and system from Taiyuan University of Technology, China in 1997. She is now working as associate professor and supervisor of the graduates with Taiyuan University of Technology. Her fields of interest include EM and microwave technology, antennas, and intelligent information processing.

Nan Li was born in China in May 1985. She now is studying in Taiyuan University of Technology for getting Master of Engineering in electronic circuit and system. Her fields of interest include antennas and intelligent information processing.

Keming Xie was born in China in April 1944. He is now working as professor and Ph.D. supervisor with Taiyuan University of Technology. His fields of interest include automatic control and intelligent information processing.

An Arbitrary-length and Multiplierless DCT Algorithm and Systolic Implementation

Zhenbing Liu, Jianguo Liu and Guoyou Wang

State Key Laboratory for Multispectral Information Processing Technologies, Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan 430074, P.R.China
liuzb0618@hotmail.com, jgliu@ieee.org, gywang@mail.hust.edu.cn

***Abstract*—Discrete Cosine transform (DCT) is an important tool in digital signal processing. In this paper, a novel algorithm to perform DCT multiplierlessly is proposed. First, by modular mapping and truncating Taylor series expansion, the DCT is expressed in the form of the product of the constants and discrete moments. Second, by performing appropriate bit operations and shift operations in binary system, the product can be transformed to some additions of integers. The proposed algorithm only involves integer additions and shifts because the discrete moments can be computed only by integer additions. An efficient and regular systolic array is designed to implement the proposed algorithm, and the complexity analysis is also given. Different to other fast Cosine transforms, our algorithm can deal with arbitrary length signals and get high precision. The approach is also applicable to multi-dimensional DCT and DCT inverses.**

***Index Terms*—discrete Cosine transform, moments, multiplierless, systolic arrays**

I. Introduction

Since the discrete cosine transform (DCT) is frequently encountered in signal and image processing applications and it is highly computation-intensive, efficient algorithms and architectures are suggested for its implementation in dedicated VLSI [1-3]. Amongst the existing VLSI systems, systolic architectures have been extensively popular not only due to the simplicity of their design based on repetitive identical processing elements (PE) with regular and local interconnections; but also for their potential of using high level of pipelining in small chip-area with low power consumption.

There are three different design styles for DCT algorithm: 1) Direct factorization of the DCT matrix [1,2].; 2) indirect computation through the fast Fourier transform (FFT)[3,4]; or through the Discrete Hartley transform (DHT)[5].; and 3) algorithms based on complexity theory[6]..

In this paper, based on our work in [7,8], we proposed a multiplierless algorithm for arbitrary length DCT: by modular mapping and truncating Taylor series expansion, the DCT is expressed in the form of the product of the constants and discrete moments; then the product can be transformed to some additions of integer into additions by shifting digits and accumulation of integers. Based on the approach to the fast calculation of moments [8], systolic arrays to perform 1-D DCT are presented, followed by a complexity analysis.

The rest of the paper is organized as follows. First we introduce the moments based DCT briefly in section 2. The multiplierless DCT algorithm is presented in section 3. Followed by complexity analysis, the systolic arrays are designed to compute DCT in section 4. Finally, we conclude our paper in section 5.

.

II. Moments Based DCT

The DCT of a sampled sequence $x(0), x(1), \cdots, x(N-1)$ is given by

$$X(k) = c_k \sum_{n=0}^{N-1} x(n) \cos \frac{\pi(2n+1)k}{2N} \qquad (1)$$

$k = 0,1,2,\cdots,N-1$

where $c_k = \begin{cases} 1/\sqrt{N} & for\ k=0 \\ 2/\sqrt{N} & otherwise. \end{cases}$

By modular mapping and making use of Taylor expansions we can transform DCT into computation involving moments:

Define:

$$S_{k,i} = \{ j \,|\cos \pi(2j+1)k/2N = \cos i\pi/2N, 0 \le j \le N-1\}$$

$$s_{k,i} = \{ j \,|\cos \pi(2j+1)k/2N = -\cos i\pi/2N, 0 \le j \le N-1\}$$

Then define:

$$x_{k,i} = \begin{cases} \sum_{i \in S_{k,r}} x(i) - \sum_{i \in s_{k,r}} x(i) & \text{if } S_{k,i} \bigcup s_{k,i} \ne \phi \\ 0 & otherwise. \end{cases} \qquad (2)$$

By direct substitution, Eq.1 can be rewritten as follows:

$$X(k) = c_k \sum_{n=0}^{N-1} x(n) \cos \frac{\pi(2n+1)k}{2N} = c_k \sum_{i=0}^{N-1} x_{k,i} \cos \frac{\pi i}{2N}$$

(3).

By the theorem of the extended law of the mean, we get:

© 2010 ACADEMY PUBLISHER
doi:10.4304/jcp.5.5.725-732

$$\cos\frac{\pi i}{2N}=\sum_{r=0}^{p}\frac{(-1)^r(\pi i/2N)^{2p}}{(2p)!}+R_i \qquad (4)$$

where R_i is the Taylor remainder term.

Substituting (4) into (3), yields:

$$X(k)=c_k x_{k,0}+c_k\sum_{i=0}^{N-1}x_{k,i}\sum_{r=0}^{p}\frac{(-1)^r(\pi i/2N)^{2r}}{(2r)!}+R_p$$

$$=c_k x_{k,0}+c_k\sum_{r=0}^{p}\sum_{i=1}^{N-1}\frac{(-1)^r\pi^{2r}}{(2N)^{2r}(2r)!}x_{k,i}i^{2r}+R_p$$

$$=c_k x_{k,0}+c_k\sum_{r=0}^{p}\frac{(-1)^r\pi^{2r}}{(2N)^{2r}(2r)!}\sum_{i=1}^{N-1}x_{k,i}i^{2r}+R_p$$

$$=c_k x_{k,0}+c_k\sum_{r=0}^{p}a_r m_{k,2r}+R_p$$

where $a_r=\dfrac{(-1)^r\pi^{2r}}{(2N)^{2r}(2r)!}$, $m_{k,2r}=\sum_{i=1}^{N-1}x_{k,i}i^{2r}$ is the so-called moments, and R_p is the sum of all the Taylor remainder term.

If R_p is ignored, $X(k)$ can be computed as follows:

$$X(k)=c_k x_{k,0}+c_k\sum_{r=0}^{p}a_r m_{k,2r} \qquad 0\le k\le N-1 \qquad (5)$$

In effect, R_p converges to zero very rapidly and uniformly as p increases, so the approximation (5) can satisfy the accuracy requirement of most applications without computing too many terms. In addition, p can be expressed approximately as $[\log_2 N/\log_2\log_2 N]$ in practical use.

Thus, the computation of $X(k)$ using the approximation of (5) establishes the relationship between DCT and moments $m_{k,2r}$.

III. MULTIPLIERLESS DCT

In (5), there is a float-point dot product of the moments with a constant vector (a_r) to compute. When N is large, (a_r) is too small to compute. We can solve this problem by transforming the product of floating-point into additions of integers by the following steps.

When $N=2^k$, multiplying $(c_k a_r)$ by $2^{g_r}=2^{\log(2N)^{2r+4}+10}$ ($r=0,1,\text{L}\ ,p$), we can get $[c_k a_r\times 2^{g_r}]<(-1)^r 86\times 2^{\log(2N)^2+1}$ which are integers and can be represented as sums of distinct powers of 2:

$$[c_k\times 2^{2\log(2N)}]=\sum_{i=0}^{t}d_i 2^{t-i}$$

$$[c_k a_0\times 2^{m_1}]=\sum_{i=0}^{t}n_{0,i}2^{t-i}$$

$$[c_k a_1\times 2^{m_2}]=\sum_{i=0}^{t}n_{1,i}2^{t-i} \qquad (6)$$

$$\text{L}$$

$$[c_k a_p\times 2^{m_p}]=\sum_{i=0}^{t}n_{p,i}2^{t-i}$$

where d_i, $n_{m,i}=0or1$ ($m=1,\text{L}\ ,p,\ i=1,\text{L}\ ,t.$).

Since $[c_k a_r\times 2^{g_r}]<(-1)^r 86\times 2^{\log(2N)^2+1}<(-1)^r 2^{\log(2N)^2+17}$, we can let $t=\log(2N)^2+17=2\log N+19$.

Then:

$$X(k)=c_k x_{k,0}+c_k\sum_{r=0}^{p}a_r m_{k,2r}$$

$$=\frac{x_{k,0}c_k\times 2^{2\log(2N)}}{2^{2\log(2n)}}+\frac{\sum_{r=0}^{p}(a_r\times 2^{\log(2N)^{2r+4}+10})(\frac{m_{k,2r}}{2^{\log(2N)^{2\ (r-1)}-10}})}{2^{\log(2N)^6+20}}$$

$$=\frac{x_{k,0}\sum_{i=0}^{t}d_i 2^{t-i}}{2^{2\log(2N)}}+\frac{\sum_{r=0}^{p}\sum_{i=0}^{t}n_{r,i}2^{t-i}m'_{k,2r}}{2^{4\log(2N)+20}}+R'_p+R''_p$$

$$=\frac{\sum_{i=0}^{t}x_{k,0}d_i 2^{t-i}}{2^{2\log(2N)}}+\frac{\sum_{i=0}^{t}(\sum_{r=0}^{p}n_{r,i}m'_{k,2r})\times 2^{t-i}}{2^{4\log(2N)+20}}+R'_p+R''_p$$

where $m'_{k,2r}=[m_{k,2r}/2^{\log N^{r-1}-10}]$, and we can prove that $R'_p<3p/(8N)$ and $R''_p<64/N$, see appendix.

It is obvious that the larger the N, the smaller the error R'_p and R''_p, which can satisfy the accuracy requirement of most applications.

Thus, $X(k)$ can be approximately computed as follows:

$$X(k)=\frac{\sum_{i=0}^{t}x_{k,0}d_i 2^{t-i}}{2^{2\log(2N)}}+\frac{\sum_{i=0}^{t}(\sum_{r=0}^{p}n_{r,i}m'_{k,2r})\times 2^{t-i}}{2^{4\log(2N)+20}} \qquad (7)$$

(6) can be computed in advance once the signal length N is determined. Since $n_{r,i}=0or1$ and $m'_{k,2r}$ can be performed by shifting digits of $m_{k,2r}$ to the left $\log N^{r-1}-10$ places in binary system, the computation of $\sum_{r=0}^{p}n_{r,i}m'_{k,2r}$ only involves accumulations of $m'_{k,2r}$.

By shifting digits of $\sum_{r=0}^{p} n_{r,i} m'_{k,2r}$ to the left $t-i$ places in binary system, the multiplication of 2^{t-i} can be performed, and shifting $\sum_{i=0}^{t}(\sum_{r=0}^{p} n_{r,i} m'_{k,2r})\times 2^{t-i}$ to the right $2^{4\log(2N)+20}$ places in binary system, $\sum_{i=0}^{t}(\sum_{r=0}^{p} n_{r,i} m'_{k,2r})\times 2^{t-i}/2^{4\log(2N)+20}$ can be performed. Using the same method, we can get $\sum_{i=0}^{t} x_{k,0} d_i 2^{t-i}/2^{2\log(2N)}$.

By doing so, all the products of floating-point constants with moments in (5) are eliminated, replaced by shifting the digits in binary system and accumulations of integers. It is obvious that less than $p(t+1)+2$ additions of integers and $t+2+p$ shifts are required to implement $X(k)$ with (7) once all moments are produced.

When $N \neq 2^k$, we can get the same result by replacing $2^{\log(2N)^{2r+4}+10}$ with $2^{[\log(2N)^{2r+4}]+10}$ in the above procedures, and the only difference is here $t=[2\log N+18]$ .For convenient, we let $t=[2\log N+18]$ either when $N=2^k$ or when $N \neq 2^k$.

Ⅳ. Complexity Analysis

According to section 2 and 3, the 1-D DCT can be implemented by the following procedures:

(Input $p=[\log_2 N/\log_2 \log_2 N]$, $t=[2\log N+18]$)

1. Compute $x_{k,i}$ and $n_{r,i}$ using (2) and (6);
2. Compute $m_{k,r}$ using the method in [7];
3. Compute $X(k)$ with (7).

Next we analyze the complexity of the algorithm. Step 1 constitutes a preprocessing step which involves equations (2) and (6). It is noteworthy that $[a_r \times 2^{g_r}]$, S_{ki} and $n_{r,i}$ are real numbers only with relation to N and can be obtained in advance, so these computations are not part of the real-time system. To compute $m_{k,r}$, we use the $2p$-network method presented in [8], which require less than $(2p+1)(p+1)(N-2)N$ integer additions. From section 3, the computation of (7) involves $(p+1)[2\log N+18]N+2N$ integer additions and $([2\log N+18]+p+2)N$ shifts.

The p-network shown in Fig.1 represents a map of transforming the vector $(1,x,x^2,...x^{p-1},x^p)$ into $(1,(1+x),(1+x)^2,...(1+x)^{p-1},(1+x)^p)$. It is denoted by F_p, i.e.,

$$F_p(1,x,x^2,...x^{p-1},x^p)$$
$$=(1,(1+x),(1+x)^2,...(1+x)^{p-1},(1+x)^p)$$

In general

$$F_p^{\,n-1}(1,x,x^2,...x^{p-1},x^p)$$
$$=(1,(n-1+x),(n-1+x)^2,...,(n-1+x)^{p-1},(n-1+x)^p)$$

By substitution,

$$F_p^{\,n-1}(1,1,1,...1,1)=(1,n,n^2,...,n^{p-1},n^p)$$
$$F_p^{\,n-1}(a,a,a,...a,a)=(a,na,n^2a,...,n^{p-1}a,n^p a).$$

Let $A_i=(a_i,a_i,...,a_i,a_i)$ which is a $p+1$ -demensional vector, then we can get

$$F_p\left(F_p\cdots F_p\left(F_p\left(F_p(\mathrm{A_n})+\mathrm{A_{n\text{-}1}}\right)+\mathrm{A_{n\text{-}2}}\right)+\cdots+\mathrm{A_2}\right)+\mathrm{A_1}$$
$$=F_p^{n-1}(\mathrm{A_n})+F_p^{n-2}(\mathrm{A_{n\text{-}1}})+F_p^{n-3}(\mathrm{A_{n\text{-}2}})+\cdots+F_p^{2}(\mathrm{A_3})+F_p(\mathrm{A_2})+\mathrm{A_1}$$
$$=\left(\sum_{i=1}^{n}a_i,\sum_{i=1}^{n}a_i i,\sum_{i=1}^{n}a_i i^2,\cdots,\sum_{i=1}^{n}a_i i^{p-1},\sum_{i=1}^{n}a_i i^p\right)$$

For emphasis, the above equation is rewritten as

$$Moment_p\left(a_n,a_{n-1},a_{n-2},\cdots,a_2,a_1\right)$$
$$=F_p\left(F_p\cdots F_p\left(F_p\left(F_p(\mathrm{A_n})+\mathrm{A_{n\text{-}1}}\right)+\mathrm{A_{n\text{-}2}}\right)+\cdots+\mathrm{A_2}\right)+\mathrm{A_1}$$
$$=\left(\sum_{i=1}^{n}a_i,\sum_{i=1}^{n}a_i i,\sum_{i=1}^{n}a_i i^2,\cdots,\sum_{i=1}^{n}a_i i^{p-1},\sum_{i=1}^{n}a_i i^p\right) \tag{8}$$

Equation (8) can be proved by mathematical induction. These components of the resultant vector are known as 1-D moments. To compute these 1-D moments, F_p is used $n-1$ times in the iteration procedure. As F_p only involves $p(p+1)/2$ additions, so $Moment_p$ only includes integer additions. The algorithm for computing 1-D moments can be described below:

```
Subroutine $Moment_p(a_n,a_{n-1},a_{n-2},\cdots,a_2,a_1)$
Input initial $p$ , $a_1,a_2,\cdots,a_{n-1},a_n$
$Moment = A_n$
for $i=1$ to $n-1$
    Perform $F_p(Moments)$
     $Moments = F_p(Moments)+A_{n-i}$
end for
```

Ⅴ. Systolic Implementation for the Proposed Algorithm

Though it seems that the proposed algorithm requires a larger number of additions than other DCTs, it is easily implemented by systolic arrays efficiently because it does not involves multiplications which is time consuming.

The general systolic network for implementing DCT consists of the moment generator and the shifting and accumulation array. The moments generator [8] formed of $N-1$ $2p$-network with a row of adder-latch could be used to generate the 1-D moments $m_{k,2r}$ and $x_{k,0}$ after receiving the input $x(0), x(1), \cdots, x(N-1)$, as in Fig.2. There are altogether $N+2p^2+2p+1$ adders in this generator. The data are input into and output from the generator in every clock cycle, and the total execution time of the computing procedure is equal to $(2p+3)N-2p+1$ clock periods. After $m_{k,2r}$ is produced, $m'_{k,2r}$ can be computed by shifting digits of $m_{k,2r}$ to the left $\log N^{r-1}-10$ places.

The shift and accumulation array receives $m'_{k,2r}$ and performs (7), as in Fig.3. Since $n_{r,i}=0 or 1$, the computation of $\sum_{r=0}^{p} n_{r,i} m'_{k,2r}$ only involves accumulations of $m'_{k,2r}$ $i = 0, 1, ..., \mathrm{p}$. It is noted that all the $m'_{k,2r}$ s ($i = 0, 1, ..., \mathrm{p}$) are produced in the same time. Then by shifting digits of $\sum_{r=0}^{p} n_{r,i} m'_{k,2r}$ to the left $t-i$ places, the multiplication of 2^{t-i} can be performed, and shifting $\sum_{i=0}^{t}(\sum_{r=0}^{p} n_{r,i} m'_{k,2r}) \times 2^{t-i}$ to the right $2^{4\log(2N)+20}$, $\sum_{i=0}^{t}(\sum_{r=0}^{p} n_{r,i} m'_{k,2r}) \times 2^{t-i} / 2^{4\log(2N)+20}$ can be performed. After $\sum_{i=0}^{t} x_{k,0} d_i 2^{t-i} / 2^{2\log(2N)}$ is implemented in the same way, we can complete (7) by adding it with $\sum_{i=0}^{t}(\sum_{r=0}^{p} n_{r,i} m'_{k,2r}) \times 2^{t-i} / 2^{4\log(2N)+20}$. This array was formed by about $pt\log(pt)$ adders and some shifters, and the execution time is about $\log(pt)$ clock periods.

To perform DCT, $N+2p^2+2p+1+pt\log(pt)$ adders and some shifter are required, and the total execution time is about $(2p+3)N-2p+1+\log(pt)$ clock cycles.

The hardware- and time-complexities of the proposed systolic realization along with those of the existing structures [9-13] are listed in Table I. T_M and T_A are, respectively, the times involved in performing one multiplication and one addition in those structures in per clock cycle. Since $p \ll N$ and $t \ll N$, the number of adders and clock periods are approximately denoted as N and $2pN$ respectively.

It seems that the proposed architectures is slower than others because the latency is $2pN$ (p is nearly a constant no more than 5) cycle periods, but during every cycle period, it only performs integer operations (the time is only T_A) while others perform floating-point multiplications which are much slower than integer additions. So the total time required in our method could be shorter than other methods.

When the hardware is concerned, about N adders in the proposed array is required, which is two times than the existing methods. But there are no multipliers required, while the existing methods need about $N/2$ multipliers. In another words, the total number of multipliers and adders in the existing methods is the same as that of adders in the proposed method. Since an adder is much simpler than a multiplier, the proposed structure is simpler and easier to design.

Besides, our array only involves integer operations, so the only error is very small, and hence our structure can get high precision without additional hardware.

VI. Conclusions

In this paper, novel algorithm to perform 1-D DCT is proposed, which does not involves multiplication operations and can deal with any-length DCT. Then, the systolic structure for the proposed algorithm is designed, followed by a complexity analysis. These results can be extended to multi-DCT and IDCT

Compared with the existing method, the new method has also the following advantages: there are no multiplications in our method, which is superior to the $O(N\log_2 N)$ in classical DCT; exponential functions have been replaced by simple polynomial functions and all multiplications have been changed into additions, which decreases the computational cost and memory requirement; it can accommodate data samples of arbitrary length and compute any portion of frequency values of DCT. In addition, since our method only involves integer operations, it produces nice accuracy and convergence property.

APPENDIX A

$$\begin{aligned}
R_p^{'} &= \sum_{r=0}^{p}((a_r \times 2^{m_r})(m_{k,2r}/2^{\log(2N)^{2(r-1)}-10}))/2^{6\log(2N)+20} - (\sum_{r=0}^{p}[a_r \times 2^{m_r}][m_{k,2r}/2^{\log(2N)^{2(r-1)}-10}])/2^{6\log(2N)+20} \\
&= \sum_{r=0}^{p}((a_r \times 2^{m_r})(m_{k,2r}/2^{\log(2N)^{2(r-1)}-10}))/2^{6\log(2N)+20} - \sum_{r=0}^{p}([a_r \times 2^{m_r}](m_{k,2r}/2^{\log(2N)^{2\ (r-1)}-10}))/2^{6\log(2N)+20} \\
&\quad + \sum_{r=0}^{p}([a_r \times 2^{m_r}](m_{k,2r}/2^{\log(2N)^{2\ (r-1)}-10}))/2^{6\log(2N)+20} - \sum_{r=0}^{p}([a_r \times 2^{m_r}][m_{k,2r}/2^{\log(2N)^{2(r-1)}-10}])/2^{6\log(2N)+20} \\
&\le \sum_{r=0}^{p}(a_r \times 2^{m_r} - [a_r \times 2^{m_r}])(m_{k,2r}/2^{\log(2N)^{2(r-1)}-10})/2^{6\log(2N)+20} \\
&\quad + \sum_{r=0}^{p}[a_r \times 2^{m_r}](m_{k,2r}/2^{\log(2N)^{2(r-1)}-10} - [m_{k,2r}/2^{\log(2N)^{2(r-1)}-10}])/2^{6\log(2N)+20} \\
&\le \sum_{r=0}^{p}(m_{k,2r}/2^{\log(2N)^{2(r-1)}-10})/2^{6\log(2N)+20} + \sum_{r=0}^{p}[a_r \times 2^{m_r}]/2^{6\log(2N)+20} \\
&< \sum_{r=0}^{p}(256 \times (2N)^{2\ (r+1)}/2^{\log(2N)^{2\ (r-1)}-10})/2^{6\log(2N)+20} + \sum_{r=0}^{p}84 \times 2^{2\log(2N)+10}/2^{6\log(2N)+20} \\
&\le \sum_{r=0}^{p}256/2^{2\log(2N)+10} + \sum_{r=0}^{p}84/2^{4\log(2N)+10} \\
&= p/2^{\log N+2} + 84p/2^{\log N+10} \\
&< p/(4N) + p/(8N) = 3p/(8N)
\end{aligned}$$

$$\begin{aligned}
R_p^{''} &= x_{k,0}[c_k \times 2^{2\log(2n)}]/2^{2\log(2n)} - x_{k,0}c_k \times 2^{2\log(2n)}/2^{2\log(2n)} \\
&= x_{k,0}([c_k \times 2^{2\log(2n)}] - c_k \times 2^{2\log(2n)})/2^{2\log(2n)} \\
&= x_{k,0}/2^{2\log(2n)} \\
&< (\max_i x_i)N/2^{2\log(2n)} \\
&< 256N/2^{2\log(2n)} = 64/N
\end{aligned}$$

ACKNOWLEDGMENT

This work was supported in part by NSFC under grants 60672060 and the Research Fund for the Doctoral Program of Higher Education

REFERENCES

[1] Z. Wang, "Fast algorithms for the discrete W transform and for the discrete Fourier transform," *IEEE Trans. Acoust.. Speech, Signal Process.*, vol. 32, pp. 803- 816, 1983.

[2] B. G. Lee, "A new algorithm for the discrete cosine transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, pp. 1243-1245, 1984.

[3] M. J. Narasimha and A. M. Peterson, "On the computation of the discrete cosine transform," *IEEE Trans. Compur.*, vol. COM-26, no. 6, pp. 934-936, 1978.

[4] M. T. Heideman, "Computation of an odd-length DCT from a realvalued DlT of the same length," *IEEE Trans. Signal Process.*, vol. 40, no. 1, pp. 54-61, Jan. 1992.

[5] H. Malvar, "Fast computation of discrete cosine transform through fast Hartley transform," *Electronic Letters*, vol. 22, no. 7, pp. 352-353, 1986.

[6] P. Yip and K. R. Rao, "Fast decimation-in-time algorithms for DST's and DCT's,"

[7] J. G. Liu, H. F. Li, F. H. Y. Chan, and F. K. Lam, "Fast Discrete Cosine Transform via Computation of Moments," *Journal of VLSI signal processing*, vol. 19, pp. 257-268, 1998.

[8] F. H. Y. Chan, F. D. Lam, H. F. Li, and J. G. Liu. "An all adder systolic structure for fast computation of moments," *J.VLSI Signal Process*, vol. 12, pp. 159-175, 1996.

[9] J.-I. Guo, C.-M. Liu, and C.-W. Jen, "A new array architecture for primelength discrete cosine transform," *IEEE Trans. Signal Processing*, vol. 41,no. 1, pp. 436–442, Jan. 1993.

[10] C. Cheng and K. K. Parhi, "A novel systolic array structure for DCT," *IEEE Trans. Circuits Syst-II: Express Briefs*, vol. 52, no. 7, pp. 366–369, July 2005.

[11] D. F. Chiper, M. N. S. Swamy, M. O. Ahmad, and T. Stouraitis, “Systolic algorithms and a memory-based design approach for a unified architecture for the computation of DCT/DST/IDCT/IDST,” *IEEE Trans. Circuits Syst-I: Regular Papers,* vol. 52, no. 6, pp. 1125–1137, June 2005.

[12] P. K. Meher, “Systolic designs for DCT using a low-complexity concurrent convolutional formulation,” *IEEE Trans. Circuits & Systems for Video Technology*, vol. 16, no. 9, pp. 1041–1050, Sept. 2006.

[13] P. K. Meher, “A new convolutional formulation of the DFT and efficient systolic implementation,” *in Proc. IEEE Int. Region 10 Conf. (TENCON’05),* Nov. 2005, pp. 1462–1466.

Zhenbing Liu was born in Shandong, China in 1980. He received the B.S. degree in math from Qufu Normal University in Qufu of Shandong Province in 2003, and M.S degree in math from Huazhong University of Science and Technology in Wuhan of Hubei province in 2006, both in china .

He is pursing his Ph.D degree at Huazhong University of Science and Technology. His research interests include parallel algorithm and structure, and svm, and have published 7 papers.

Jianguo Liu was born in Hubei, china in 1952. He received his B.S. degree in math from Wuhan University of Technology in 1983, M.S degree in math from Huazhong University of Science and Technology in 1984, both in Wuhan, Hubei province of china, and Ph.D degree in computer science from Hong Kong University in 1996 in Hongkong.

He was a visiting scholar with the Medical Image Processing Group, Department of Radiology, University of Pennsylvania, Philadelphia, from December 1998 to July 2004. He is currently Professor and Deputy Director of the Key Laboratory of the State Education Ministry for Image Processing and Intelligent Control, Huazhong University of Science and Technology in Wuhan. His interests include signal processing, image processing, parallel algorithm and structure, and pattern recognition.

Prof. Liu is a member of IEEE.

Guoyou Wang received the B.S. and M.S. degree in computer science from Huazhong University of Science and technology in Wuhan of Hubei province in China in 1988 and 1992 respectively.

Prof. Wang is a Professor with Huazhong University of Science and technology. His research interests include image compression and data mining.

He is a member of IEEE.

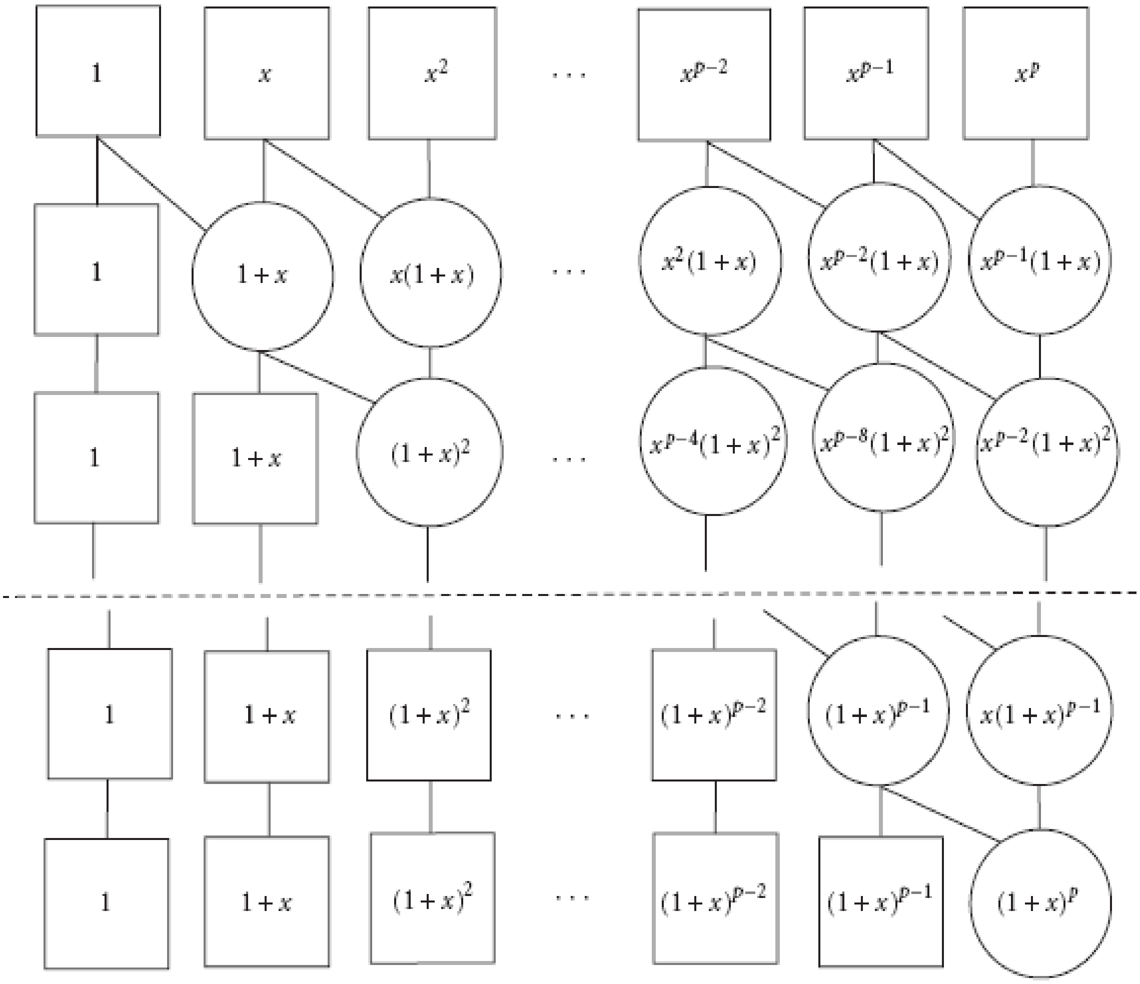


Fig.1 the p-network

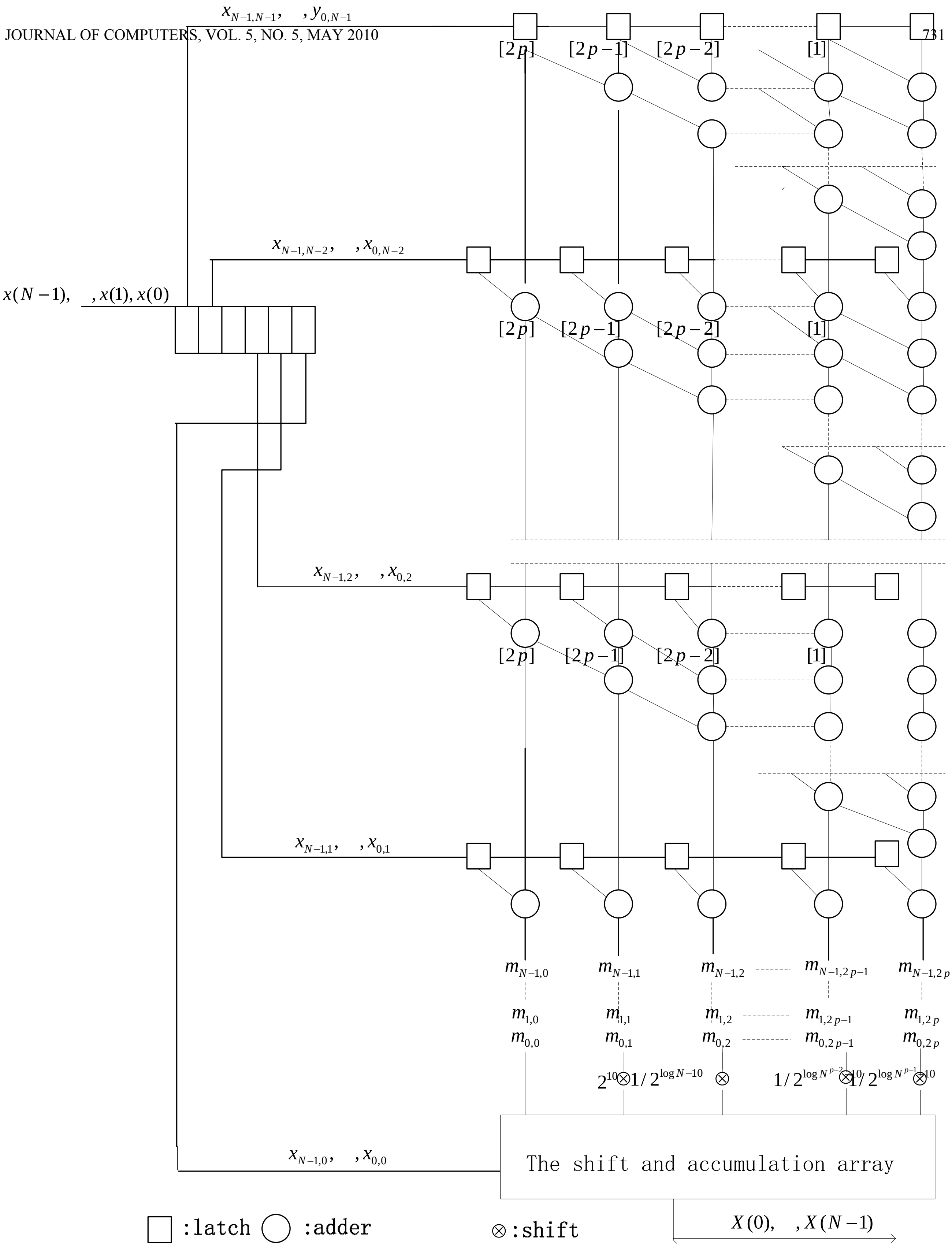

$x_{N-1,N-1}, \quad , y_{0,N-1}$
[2p]
[2p−1]
[2p−2]
[1]
$x_{N-1,N-2}, \quad , x_{0,N-2}$
$x(N-1), \quad , x(1), x(0)$
$x_{N-1,2}, \quad , x_{0,2}$
$x_{N-1,1}, \quad , x_{0,1}$
$m_{N-1,0}$
$m_{N-1,1}$
$m_{N-1,2}$
$m_{N-1,2p-1}$
$m_{N-1,2p}$
$m_{1,0}$
$m_{1,1}$
$m_{1,2}$
$m_{1,2p-1}$
$m_{1,2p}$
$m_{0,0}$
$m_{0,1}$
$m_{0,2}$
$m_{0,2p-1}$
$m_{0,2p}$
$2^{10}\otimes 1/2^{\log N-10} \otimes$
$1/2^{\log N^{p-2}-10}\otimes 1/2^{\log N^{p-1}-10}\otimes$
$x_{N-1,0}, \quad , x_{0,0}$
The shift and accumulation array
$X(0), \quad , X(N-1)$
□ :latch ○ :adder
⊗:shift

Fig.2 The systolic array for 1-D DCT

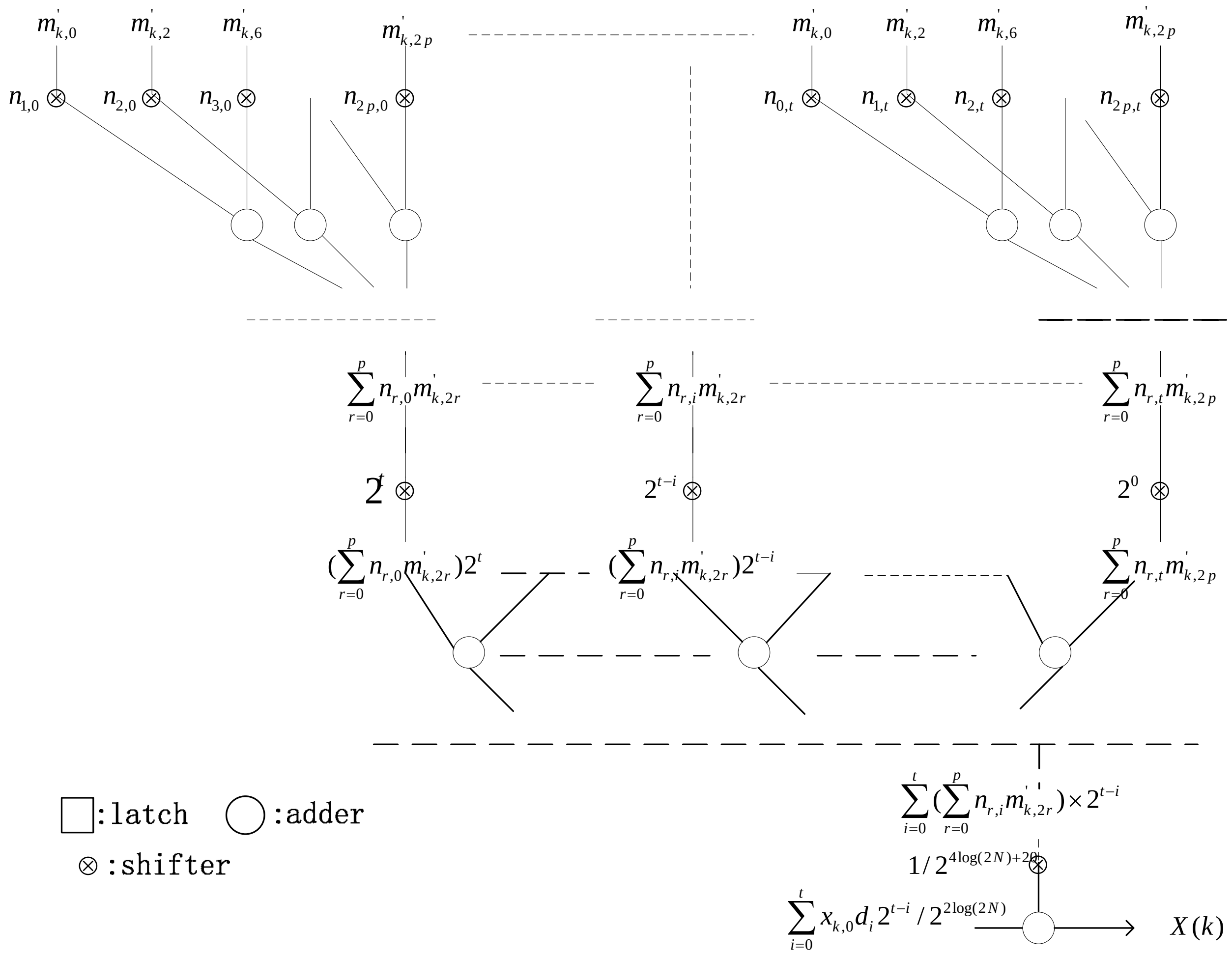

$m'_{k,0}$
$m'_{k,2}$
$m'_{k,6}$
$m'_{k,2p}$
$n_{1,0}$
$n_{2,0}$
$n_{3,0}$
$n_{2p,0}$
$n_{0,t}$
$n_{1,t}$
$n_{2,t}$
$n_{2p,t}$
$\sum_{r=0}^{p} n_{r,0} m'_{k,2r}$
$\sum_{r=0}^{p} n_{r,i} m'_{k,2r}$
$\sum_{r=0}^{p} n_{r,t} m'_{k,2p}$
2^t
2^{t-i}
2^0
$(\sum_{r=0}^{p} n_{r,0} m'_{k,2r})2^t$
$(\sum_{r=0}^{p} n_{r,i} m'_{k,2r})2^{t-i}$
$\sum_{r=0}^{p} n_{r,t} m'_{k,2p}$
□:latch ○:adder
⊗:shifter
$\sum_{i=0}^{t}(\sum_{r=0}^{p} n_{r,i} m'_{k,2r}) \times 2^{t-i}$
$1/2^{4\log(2N)+2Q}$
$\sum_{i=0}^{t} x_{k,0} d_i 2^{t-i} / 2^{2\log(2N)}$
$X(k)$

Fig.3 The shift and accumulation array

TABLE I.
HARDWARE- AND TIME-COMPLEXITIES OF VARIOUS STOLIC STRUCTURES FOR THE DCT

Structures	cycle-time	Multipliers	adders	Cycles
Guo *et al.* [9]	$T_M + T_A$	$(N+1)/2$	$(N+3)/2$	$N-1$
Chiper *et al.* [11]	$T_M + T_A$	$N-1$	$N+1$	$(N-1)/2$
Cheng *et al.* [10]	$T_M + 3T_A$	$(N-1)/2$	$3(N-1)/2$	$(N-1)/2$
Meher [12]	$T_M + T_A$	$N/2+3$	$N/2+5$	$N/2-1$
Meher [13]	$2(T_M + T_A)$	$N/2-1$	$N/2+9$	$N/4-1$
Proposed	T_A	0	N	$2pN$

Preserving Private Knowledge In Decision Tree Learning

Weiwei Fang
Information Engineering School, University of Science and Technology Beijing
Computer Center, Beijing Information Science and Technology University, China
Email: Liveinbetter@163.com

Bingru Yang and Dingli Song
Information Engineering School, University of Science and Technology Beijing
Email: sdlhr617@sohu.com

***Abstract*—Data mining over multiple data sources has become an important practical problem with applications in different areas. Although the data sources are willing to mine the union of their data, they don't want to reveal any sensitive and private information to other sources due to competition or legal concerns. In this paper, we consider two scenarios where data are vertically or horizontally partitioned over more than two parties. We focus on the classification problem, and present novel privacy preserving decision tree learning methods. Theoretical analysis and experiment results show that these methods can provide good capability of privacy preserving, accuracy and efficiency.**

***Index Terms*—Privacy Preserving, Data Mining, Decision Tree, Homomorphic encryption**

I. Introduction

In present, great advances in networking and databases technologies make it easy to distribute data across multi parties and collect data on a large scale for sharing information. Distributed data mining such as association rule mining and decision tree learning are widely used by global enterprises to obtain accurate market underlying information for their business decision. Although different enterprises are willing to collaborate with each other to data mine on the union of their data, due to legal constraints or competition among enterprises, they don't want to reveal their sensitive and private information to others during the data mining process.

There has been growing concern that use the technology of gaining knowledge from vast quantities of data is violating individual privacy. This has lead to a backlash against this technology. For example, Data-Mining Moratorium Act introduced in the U.S Senate that would have banned all data mining programs by the U.S. Department of Defense. Privacy preserving data mining (PPDM) has emerged to address this problem, and become a challenging research area in the field of data mining (DM) and knowledge discovery (KD). The main goal of preserving privacy data mining is to enable such win-win-win situations: The knowledge present in the data is extracted for use, the individual's privacy is protected, and the data holder is protected against misuse or disclosure of the data. [1].

The method of preserving privacy data mining depend on the data mining task (i.e., association rule, classification, clustering, etc.) and the data sources distribution manner (i.e., *centralize* where all transactions are stored in only one party; *horizontally* where every involving party has only a subset of transaction records, but every record contains all attributes; *vertically* where every involving party has the same numbers of transaction records, but every record contains partial attributes). In this paper, we particularly focus on applying preserving privacy data mining methods on the decision tree learning over vertically and horizontally partitioned data.

Supported by the Natural Science Foundation of China under Grant No. 69835001, Scientific Research Common Program of Beijing Municipal Commission of Education KM200811232013.

doi:10.4304/jcp.5.5.733-740

The rest of the paper is organized as follows. In the next section, we will introduce the related work in preserving privacy data mining and the contribution we did. In section 3, we provide some background technologies such as distributed decision tree learning, some secure multiparty computation and Homomorphic encryption. In section 4, we present our work of how to build a distributed decision tree over vertically partitioned data, which doesn't reveal privacy during the stages of building and classification. In section 5, we present our work of how to build a distributed decision tree over horizontally partitioned data. Section 6 shows the experimental results and privacy analysis. Conclusion is given at the last section.

II. Related work

Preserving privacy data mining provides methods that can compute or approximate the output of a data-mining algorithm without revealing at least part of the sensitive information about the data. Generally speaking, there are two approaches in privacy preserving data mining. One is using randomization techniques [2,3,4], that is, adding "noise" to the data before the data mining process, and using techniques that mitigate impact of the noise from the data mining results, however, recently there has been much debate about this kind of method, e.g., accuracy loss of mining results as altering the original data, inference problem can be derived from the reconstruction model, etc.

The other approach is using secure multiparty computation (SMC) techniques, such as secure sum, secure set union, secure size of intersection and scalar product, etc. In [6], Clifton has proposed to apply secure scalar product methods on association rules over horizontally and vertically partitioned data, respectively. In [7], Vaidya proposed algorithms on building decision tree, however, the tree on each party doesn't contain any information that belongs to other party, the drawback of this method is that the resulting class can be altered by a malicious party.

The contributions in this paper are as follows:

1) Methods proposed in this paper can be used in two contexts: vertically and horizontally partitioned data;

2) In the context of vertically partitioned data, as we apply a new classifying model, the private information can be preserved not only in the stage of building tree, but also in the classification stage;

3) In the context of horizontally partitioned data, we apply a new technology homomorphic encryption, which hasn't been used in this field; each party only can obtain the mining knowledge without leaking their own private information;

4) Both of these methods can be applied on more than two parties.

III. Background

A. Decision Tree Learning s

A decision-tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. Generally speaking, the basic algorithm for decision-tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. The presentation here is rather simplistic and very brief and we refer readers to Ref. [1] for an in-depth treatment of the subject.

Obviously, the key of decision-tree induction is selecting the attribute that will best separate the samples into individual classes, for it plays an importance in not only the effectiveness of induction, but also the quality of mining rules.

B. secure multiparty computation

Secure multiparty computation (SMC) is the problem of evaluating a function of two or more parties' secret inputs, such that each party finally hold a share of the function output and no more else is revealed, except what is implied by the party's own inputs and outputs. SMC problem was firstly introduced by Yao and extend by Goldreich and others. These works use a similar methodology: the function f to be computed is represented as a Boolean circuit, and then the parties run a protocol for every gate in the circuit. Every participant gets shares of the input wires and the output wires for every gate. Since determining which share goes to which party is done randomly, a party's own share tells it nothing. Upon completion, the parties exchange their shares, enabling each to compute the final result.

In this paper, we proposed a PPDM method by applying PCIWL (Protocol for Comparing Information Without Leaking) and MNP (Mix Network Protocol), both of which belong to issues of SMC technology. We encourage readers who want deep understanding of the above two techniques to start with Ref. [8].

C. Homomorphic encryption

Computation on encrypted data does not make sense unless the encryption transformation being used has some homomorphic properties. The homomorphic encryption presented in this paper is based on a concept named privacy homomorphism [2], which was formally introduced by Rivest in 1978 as a tool for processing encrypted data. Basically, they are encryption functions E: T→T' which allow to perform a set F' of operations on ciphertext without knowledge of the decryption function D. Knowledge of D allows to recover the outcome of the corresponding set F of operations on plaintext. The security gain is especially apparent in multilevel security scenarios. That is, sensitive data will be encrypted by the classified institute, be processed by the unclassified contractor, and the result be decrypted by the classified institute [3].

Let S be a set, and S’ be a possibly different set with the same cardinality as S. Let D: S→S’ be bijective. D is decryption function, and the encryption function is E. Assign an algebraic system for plaintext operations by:

Let S be a set, and S’ be a possibly different set with the same cardinality as S. Let D: S→S’ be bijective. D is decryption function, and the encryption function is E. Assign an algebraic system for plaintext operations by:

$$U=<S;f_1,\cdots,f_k; P_1,\cdots, P_m; s_1,\cdots, s_n>$$

Where the fi is operator, the Pi is predicate, and the si is distinct constant. Assign converse computation of U with encrypted data by:

$$C=<S';f'_1,\cdots,f'_k; P'_1,\cdots, P'_m; s'_1,\cdots, s'_n>$$

Where the f’i, P’i and s’i are the encrypted version of fi, Pi and si respectively. The mapping D is called a privacy homomorphism if it satisfies the following conditions:

1)
$$\forall i(a,b,c,...)(f'_i(a,b,...)=c$$
$$\Rightarrow f_i(D(a),D(b),...)=D(c));$$

2)
$$\forall i(a,b,c,...)(P'_i(a,b,...)\equiv P_i(D(a),D(b),...));$$

3) $D(s'_i)= s_i$

In order for C and D to be of any use as a protection, the following additional constraints should be satisfied:

1) D and E are easy to compute;

2) The functions f’i and predicates P’i in C are efficiently computable;

3) E is a non-expanding cipher or an expanding cipher whose crypto text has a representation only marginally larger that the corresponding plaintext;

4) The operations and predicates in C should not be sufficient to yield an efficient computation of D.

IV. Privacy Preserving Decision Tree Learning over Vertically Partitioned Data

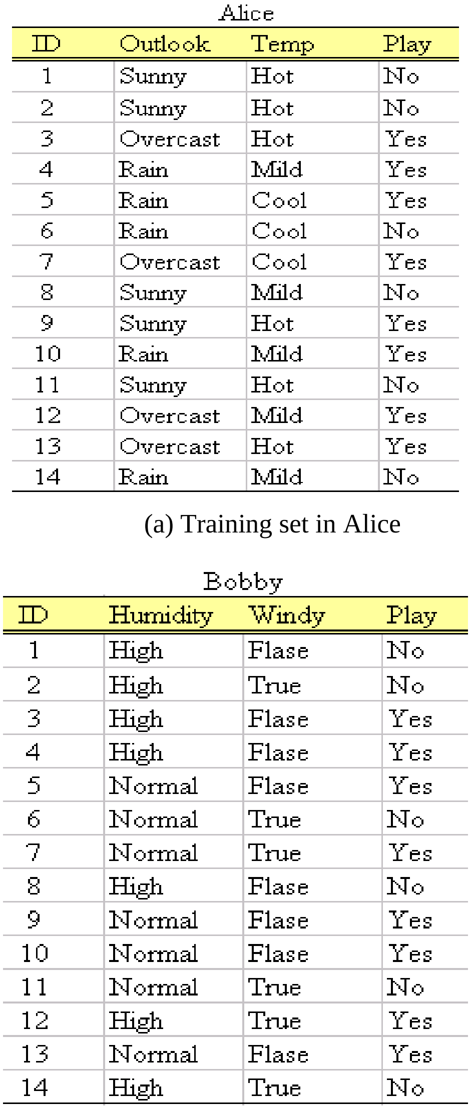
In this section we address the issue of privacy preserving distributed decision-tree mining over vertically partitioned data. Specifically, we consider a scenario in which two or more parties owning confidential databases wish to run a data-mining algorithm on the union of their databases, without revealing any original information. We propose a privacy preserving distributed decision tree learning method based on ID3 [1], which is applied in mining concentrative database and uses information entropy to choose the best prediction attribute.

Privacy preserving can mean many things [5]: Protecting specific individual values, breaking the link between values and the individual they apply to, protecting source, etc. This paper aims for a high standard of privacy: Not only individual entities are protected, but to the extent feasible even the schema (attributes and possible attribute values) are protected from disclosure.

A. Tree building

Let R be the set of condition attributes and C be the class attribute, we make assumptions that the database is vertically partitioned between k parties; each party Pi only knows its own attributes Ri, transaction ID and attribute C are known to all parties.

We take an example, as figure 1 shows, the class attribute is play, which is determined by four condition attributes, such as outlook, temp (possessed by Alice) and humidity, windy (possessed by Bobby).



Alice

ID	Outlook	Temp	Play
1	Sunny	Hot	No
2	Sunny	Hot	No
3	Overcast	Hot	Yes
4	Rain	Mild	Yes
5	Rain	Cool	Yes
6	Rain	Cool	No
7	Overcast	Cool	Yes
8	Sunny	Mild	No
9	Sunny	Hot	Yes
10	Rain	Mild	Yes
11	Sunny	Hot	No
12	Overcast	Mild	Yes
13	Overcast	Hot	Yes
14	Rain	Mild	No

(a) Training set in Alice

Bobby

ID	Humidity	Windy	Play
1	High	Flase	No
2	High	True	No
3	High	Flase	Yes
4	High	Flase	Yes
5	Normal	Flase	Yes
6	Normal	True	No
7	Normal	True	Yes
8	High	Flase	No
9	Normal	Flase	Yes
10	Normal	Flase	Yes
11	Normal	True	No
12	High	True	Yes
13	Normal	Flase	Yes
14	High	True	No

(b) Training set in Bobby

Figure1 Training set

If we use the traditional ID3 algorithm to mine on the union of datasets, we can obtain the public decision tree shown in figure 2, while each party's private information are all revealed.

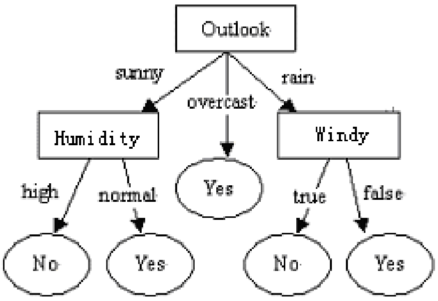


Figure 2 Public decision tree

In order to preserve each party's private data, we introduce two new notions. One is Privacy-Preserving Decision Tree, as figure 3 shows, which is stored at the miner site. The semi-honest miner only knows the basic structure of the tree, (e.g., the number of the branches at each node, the depth of each sub-tree) and which site is responsible for the decision made at each node (i.e., only know which site possesses the attribute to make decision at the node, while without the knowledge of which attribute it is and what attribute values it has); the other is Constrain Set, e.g. {AR1, BR1}, it means that this path which is form the root node to the present node (the node with the value of BR1) has determined by those attributes in the Constrain Set. When beginning to build tree, all parties will send the numbers of local attribute to miner, and the Constrain Set is initialized as {}, as Constrain Set of the present node becomes full, i.e. {AR1, BR1, AR2, BR2}, it means R is empty [1], the next node should be leaf node, which with the class attribute value c assigned to most transactions with the certain transaction IDs.

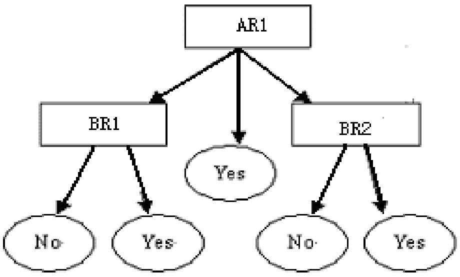


Figure 3 Privacy-preserving decision tree

Now we'll describe how it can be built and used to classify testing sets. When the miner creates a root node, it sends signal to all parties. Each party obtains the local best prediction attribute Ri by information gain measurement, then sends the attribute serial number Ri and information entropy to the miner by PCIWL (Protocol for Comparing Information Without Leaking), which ensures that no original information would be revealed at miner site or any other parties. The miner applies PCIWL to get the maximum as the global best prediction attribution, while he doesn't know the which attribute it is and what attribute values it has, he just has the knowledge that which site possesses that attribute and its' serial number, e.g., as it is shown in figure 3, the minor creates a root node AR1, which means Alice has the information at that node, and the first attribute possessed by Alice is the best prediction attribution. At the same time, the minor set {AR1} as Constraint Set of the present node.

When creating the next node, whether it's a leaf or internal node, the process is as following: Firstly, the miner sends token signal to the target party, which has possesses the best prediction attribute of previous node. Secondly, the target party receives token message, if the token signal is 0, which means Constraint Set is full, it only needs to compute the class attribute value c assigned to most transactions with the certain IDs, and send c to miner site; if the token signal is 1, which means that R isn't empty, it firstly needs to judge if all the transactions with the certain IDs have the same class attribute c, if so, then sends c to miner site; otherwise works out the intersection of transactions used previously and transactions with best prediction attribute value, and sends IDs to other parties by MNP (Mix Network Protocol), by which it guarantees that the communication process is anonymous. Thirdly, all parties compute information entropy of the local attribute corresponding to the certain IDs, and send the information entropy to the miner site by PCIWL. Finally, if the minor only receives attribute c from token party, it creates a leaf node with the value of c; if the miner receives information entropy form all parties, it chooses the maximum as the best prediction attribute, adds the attribute tag to Constant Set, and sends token to the next target party, which possesses the best prediction attribute of the present node.

B. Privacy-preserving algorithm

Assume that there are three parties named A, B and C, which respectively has ra, rb and rc condition attributes, and wants to collaboratively mining decision-tree. As the main idea we presented above, the algorithms, which comprise three parts, are as follows:

Local mining algorithm (performed by parties with token):

Input: Local training samples, token.

Output: Sending class attribute distribution to miner site, or sending IDs to other parties and information entropy to miner.

1) If token=0, computes the class attribute value c assigned to most transactions with the certain IDs, and sends c to miner site;
2) If token=1, judges if all the transactions with the certain IDs have the same class attribute c, if so, sends c to miner site;

3) If not, works out the intersection of transactions used previously and transactions with best prediction attribute value, sends IDs to other parties by MNP, and do step 4;

4) Computes information entropy and sends it to the miner site by PCIWL;

Local mining algorithm (performed by parties without token):

Input: Local training samples, transaction IDs.

Output: Sending information entropy to miner.

1) Receives transaction IDs form the token party, then computes intersection of IDs received and IDs used previously;

2) Computes information entropy corresponding to the certain IDs, and sends it to the miner site by PCIWL.

Miner site algorithm (performed by miner):

Input: Class attribute distribution from token party, or information entropy from all parties.

Output: Creating node, updating Constraint Set, sending token signal to target party.

1) If the receiving message is class attribute c from token party, creates a leaf node with the value c;

2) If the receiving message is information entropy from all parties, applies PCIWL to obtain maximum, and do step 3;

3) Creates an internal node with the value of target party's name and serial number of the best prediction attribute, adds the attribute to Constraint Set, and do step 4;

4) If Constraint Set is full, sends token=0 to the target party; otherwise sends token=1.

V. Privacy Preserving Decision Tree Learning over Horizontally Partitioned Data

In this section we address the issue of privacy preserving distributed decision-tree mining over horizontally partitioned data. Specifically, we consider a scenario in which two or more parties owning confidential databases wish to run a data-mining algorithm on the union of their databases, without revealing any original information. Our work is motivated by the need to both protect privileged information and enable its use for research or other purposes.

A. Architecture

The new architecture is depicted in Figure 4, in which HE means homomorphic encryption. It's composed of the participating databases, a calculator and a miner, which uses the following basic assumptions:

1) The database is horizontally partitioned between N participants, and there is no communication between all the participants themselves;

2) The calculator only performs auxiliary computations, without knowing their meaning and having no part of the databases;

3) Although the miner manages the data mining process and reports the results to the participants, it has no part of the databases;

4) The model is semi-honest. That is, every adversary correctly follows the protocol specification, yet attempts to learn additional information by analyzing the transcript of messages received during the execution.

5) There is no external knowledge present at any side.

There are three steps when every time selecting the best predicting attribute. Firstly, every party calculates its' local Gini index, encrypts it and then transfers the encrypted data to the calculator; secondly, the calculator computes sums of certain encrypted data without knows its' meaning and transfers it to the miner; thirdly, the miner uses secret key to decrypt them, selects the best predicting attribute and broadcasts to every party.

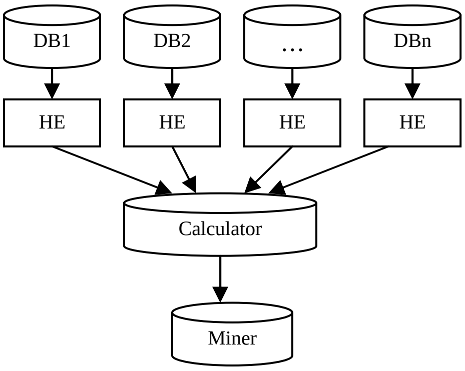


Figure4. Architecture

B. Homomorphic encryption and decryption scheme

Homomorphic encryption ensures that the computation result on two or more encrypted values is exactly the same as the encrypted result of the same computation on two or more unencrypted values.

In this paper, we proposed an additively homomorphic encryption and decryption scheme, which is as follows:

Encryption:

1) The algorithm uses a large number r, such that r= p ×q, where p and q are large security prime numbers.

2) Given x, which is a plaintext message, we compute the encrypted value y=Ep(x)=mod((x+p),r), where mod() is a common modul-operation.

Decryption:

Given y, which is a ciphertext message, we use the security key p to recover plaintext x=Dp(y)=mod(y,p).

Theorem: According to the above encryption and decryption algorithms, for every plaintext x, y and z, D (E (x)+ E (y)+ E (z)) = = x + y + z

Proof: D (E (x)+ E (y)+ E (z)) = D ((mod (x + p), n)+ (mod (y + p), n)+ (mod (z + p), n))

= D (mod (x+y+z+3p,n)) =mod (mod (x+y+z+3p,n), p)

= Mod ((mod (x + y + z), n) +mod (3p,n)), p)

= Mod ((mod (x + y + z), n), p) +mod (mod (3p,n), p)

we know n = p×q, so mod(mod(3p,n),p) =0；

As the Decryption definition says,

D(E(x)) = mod((mod(x + p),n),p) = x，

We can proof mod((mod(x + y + z),n),p) =x + y + z；

So D (E (x)+ E (y)+ E (z)) == x + y + z.

In the architecture we proposed earlier, as the calculator only sees ciphertext, has not access to the security key p, according to Domingo-Ferrer [2], with only ciphertext, it is a NP-hard problem for attacker to find the original values.

C. *Privacy-preserving algorithm*

Assume that there are three parties named A, B and C, which respectively has ma, mb and mc sample records, and wants to jointly mining decision-tree rules. As the architecture we presented above, the algorithms, which are composed of three parts, are as follows:

Local mining algorithm (performed by parties):

Input: Local training samples.

Output: Encrypted Gini Indexes of correlative attributes.

1) Scan training samples, and select sample set Sa and attribute set Ta ={Ta1, Ta2 , , , Tan} which is correlative to the present node;
2) According to definition of Gini Index [1], compute every Gini Index of attribute in set Ta, by using the formula GiniTai(Sa) = ｜Sa1｜Gini(Sa1) / ｜Sa｜+ ｜Sa2｜Gini(Sa2)/ ｜Sa｜（n≥i≥1）
3) For every element Tai (n≥i≥1)in the attribute set Ta, encrypt ma×GiniTai(Sa) (while mb×GiniTbi(Sb) for party B and mc×GiniTci(Sc) for party C) by the homomorphic encryption scheme presented above, thus get array {T’a1, T’a2 , , , T’an};
4) Send those n encrypted data to calculator;
5) Receive the best predicting attribute message Tk from miner;
6) Construct a new node’ s branches according to Tk ;

Calculate algorithm (performed by calculator):

Input: Three groups of array, which respectively contains n data.

Output: An array contained n data.

1) Receive three groups of array from three parties, which are array {T’a1, T’a2 , , , T’an}from party A, array {T’b1, T’b2 , , , T’bn} from party B and array {T’c1, T’c2 , , , T’cn } from party C;
2) Calculate each sum according to the point number in the array, T’i=T’ai+T’bi+T’ci (n≥i≥1);
3) Send an array {T’1, T’2 , , , T’n} to the miner;

Mining algorithm (performed by miner):

Input: An array {T’1, T’2 , , , T’n }.

Output: The best predicting attribute Tk.

1) Receive an array{T’1, T’2 , , , T’n}from the calculator;
2) Decrypt each T’i (n≥i≥1)by using security key p and the homomorphic decryption scheme presented above, thus get array{T1, T2 , , , Tn}, in which each element Ti(n≥i≥1)means a global computing result, that is, ma×GiniTa(Sa) + mb×GiniTb(Sb) + mc×GiniTc(Sc);
3) According to definition of Gini Index [1], select the minimum Tk from the decrypted array, which denotes the best predicting attribute;
4) Send the best predicting attribute message Tk to each party.

VI. EXPERIMENT RESULT

The experiment was conducted with Pentium IV3.2 GHz PC with 2GB memory on the Linux platform, and all algorithms were implemented in C/C++. We used the anonymous Web usage data of the Microsoft web site, which was created by sampling and processing the www.microsoft.com logs and donated to the Machine Learning Data Repository stored at University of California at Irvine Web Site.

We designed two sets of experiments. The first set is used to validate the effectiveness of preserving privacy algorithm on horizontally partitioned data; the second set is used to validate the effectiveness of preserving privacy algorithm on vertically partitioned data. In our experiments, we use 80% of the records as the training data and the other 20% as the testing data. We use the training data to build the decision trees, and then use the testing data to measure how accurate these trees can predict the class labels. The percentage of the correct predictions is the accuracy value in our figures. We repeat each experiment for multiple times, and each time the disguised data set is randomly generated from the same original data set.

As the resemblance of these two sets of experiments, in the following, we just show the first experiment results as an example. Figure 5 is shown that mining quality between non-privacy preserving approach and privacy-preserving approach in distributed decision-tree mining; Figure 6 is shown that privacy quality of privacy-preserving approach.

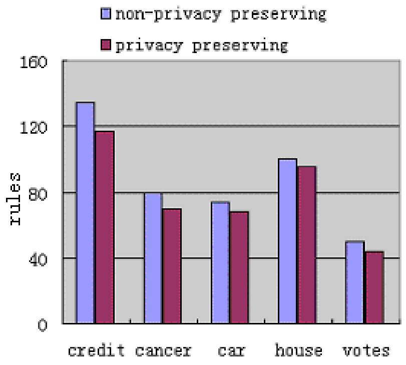


Figure5 Mining quality comparison

Figure 5 shows that compared with traditional non-privacy preserving approach, percentage of rules mined by privacy-preserving approach is at least 85%, which means although we apply privacy-preserving methods, most of the rules can also be mined; Figure 6 presents the privacy-preserving percentage is at least 82%. Experimental results show that the privacy-preserving approach we proposed can provide good capability of privacy quality without sacrificing accuracy.

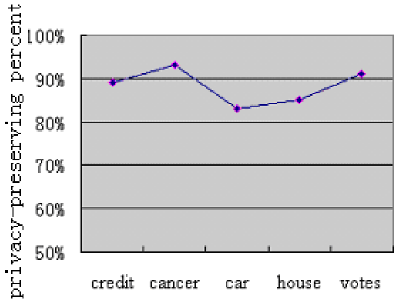


Figure 6 Privacy quality

Form the viewpoint of theoretical analysis, in the context of vertically partitioned data, when we build tree, the control is passing from site to site, except token party has the knowledge of best prediction attribute of the present node, other party even the miner doesn't know any relevant information. When we classify, the miner only knows the path of classifying process, i.e., which site handles the classifying in every step, while the information of which attribute is used to classify and values of transaction records in every party is protected. In the context of horizontally partitioned data, all row data information are encrypted, the information sent to computing center and miner are not the original information, so each party's private information are protected, and as encryption function and decryption function satisfy with Homomorphic encryption, the mining knowledge is consistent with the real results. Based on theoretical analysis and experimental results, we can conclude that methods proposed in the paper are effective.

VII. Conclusion

In this paper, we presented two privacy-preserving distributed decision-tree mining algorithms, one of which used over partitioned data is based on idea of privacy-preserving decision tree and passing control from site to site, the other of which used over horizontally partitioned data is based on the idea of homomorphic encryption. Our experimental results show that they have good capability of privacy preserving, accuracy and efficiency.

For future research, we will investigate the possibility of developing more effective and efficient algorithms. We also plan to extend our research to other tasks of data mining, like clustering and association rule, etc.

Acknowledgment

The authors wish to thank both the workmates in Data Engineering Institute, Beijing Science and Technology University and in Computer Center, Beijing Information Science and Technology University. This work was supported by the Natural Science Foundation of China under Grant No. 69835001and Scientific Research Common Program of Beijing Municipal Commission of Education KM200811232013.

References

[1] JIAWEI HAN, KAMBR M. Data mining concepts and techniques [M]. Beijing: Higher Education Press, 2001. 232 - 233.

[2] Rizvi S J, Haritsa J R. Maintaining data privacy in association rule mining [A]. In Proceedings of the 28th International Conference on Very Large Databases. Hong Kong, 2002: 682 - 693.

[3] Agrawal R, Srikant R. Privacy - preserving data mining [A]. Proceedings of the 2000 ACM SIGMOD international conference on Management of data. United States, 2000:439 - 450.

[4] Verykios V, Bertino E. State-of-the-art in Privacy preserving Data Mining,SIGMOD, 2004,33 (1).

[5] Cliffton C, Kantarcioglu M, Vaidya J. Tools for privacy preserving distributed data mining [J]. ACM SIGKDD Explorations Newsletter ,2004 ,4 (2) :28 - 34.

[6] M.Kantarcioglous and C. Clifton. Privacy preserving distributed mining of association rules on horizontally partitioned data, ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (2002), 24-31.

[7] J.Shrikant Vaidya, Privacy preserving data mining over vertically partitioned data, PH.D Thesis of Purdue University, August 2004, 28-34.

[8] Pinkas B. Cryptographic techniques for privacy-preserving data mining [J]. ACM SIGKDD Explorations Newsletter ,2006 ,4 (2) :12 - 19.

[9] Z.Yang, S.Zhong. Anonymity-preserving data collection. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, USA, August 21-24 2007

[10] S.L.Warner. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60(309): 63-69, March 2004.

[11] Bingru yang,KDK Based Double-Basis Fusion Mechanism and Its Structural Model, International Journal of Artificial Intelligence Tools, 14(3), 2005： 399-423.

[12] Piatetsky-shapiro G, Matheus C J, Knowledge Discovery Work-bench for Exploring Business Databases. International Journal of Intelligent Systems, 1992,7:675-68.

[13] Yoon J P, Kerschberg L, A Frame work for Knowledge Discovery and Evolution in Databases. IEEE Transactions on Knowledge and Data Engineering, 1993,5:973-979.

[14] Yang Bing-ru, Sun Hai-hong, Xiong Fan-lun. Mining Quantitative Association Rules With Standard SQL Queries and Its Evaluation, Journal of Computer Research and Development, 39(3), 2002: 307-312.

[15] R. Agrawal, T. Imielinski, A. Swami: "Mining Associations between Sets of Items in Massive Databases", Proc. of the ACM-SIGMOD 1993 Int'l Conference on Management of Data, Washington D.C., May 1993, 207-216.

[16] Stephen Warshall, A Theorem on Boolean Matrices, Journal of the ACM, v.9 n.1, p.11-12, Jan. 1962.

[17] Yang Bingru, Zhou Ying. The Inner Mechanism of Knowledge Discovery System and Its Influence to KDD Mainstream Development,IC-AI'02, Las Vegas,USA.

[18] Yang Bingru.A Driving Force of Knowledge Discovery in Database Main Stream — Double Bases Cooperating Mechanism, IC-AI '02, Las Vegas, USA.

[19] Renato Coppi. A Theoretical Framework for Data Mining: the "Informational Paradigm". Computational Statistics & Data Analysis, 38(2002): 501-515.

[20] D. Hand, H. Mannila, P. Smyth, Principles of Data Mining, MIT Press, Cambridge, CA, 2001.

[21] I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, San Francisco, CA, 2000.

[22] Indranil Bose, Radha K. Mahapatra. Business Data Mining—A Machine Learning Perspective. Information & Management, 39 (2001): 211-225.

[23] M.S. Chen, J. Han, and P.S. Yu. Data Mining: An Overview from a Database Perspective. IEEE Transactions on Knowledge and Data Engineering, 8(6) (1996): 866-883.

Weiwei Fang

Tutor in Beijing Information Science and Technology University; PH.D Candidate in Beijing Science and Technology University; major research field is data mining.

She has worked for morn than five years in university, mainly took lectures relevant to computer, such as programming language, database, network and etc. In 2007, she began to study in Beijing Science and Technology University as a PH.D Candidate, worked in Data Mining Institute until now. During the past two years, she has published many papers in national publication such as Computer Science, Computer Application Research, The 27th Chinese Control Conference, International Computer Science and Software Engineering Conference, etc.

Recently she has presided two projects, one is Scientific Research Common Program of Beijing Municipal Commission of Education KM200811232013, and the other is Beijing Science and Technology University Foundation 2008. She also took part in projects in Data Mining Institute, such as Natural Science Foundation of China under Grant No. 69835001, National Natural Science Foundation of China under Grant No.60875029, and etc.

Prof. Bingru Yang

He currently serves in University of Science and Technology Beijing as a chief professor, Ph.D. supervisor of School of Information Engineering and dean of Institute of Knowledge Engineering.

Prof. Dingli Song

He works as a professor in Tangshan University, teaches computer relevant lessons for more than thirty years; now he is a PH.D Candidate of University of Science and Technology Beijing, his main research field is multi-relationship data mining.

The Parameter’s Effect on the Stability in Microbial Growth Model

Hong Men, Yuhong Li,
School of Automation Engineering, Northeast Dianli University, Jilin, China
Email: menhong_bme@163.com

Fangbao Tian
Department of Modern Mechanics, University of Science and Technology of China, Hefei, China
Email: onetfbao@gmail.com

***Abstract*—A nonlinear dynamic growth model of microorganisms is developed by considering the chemotaxis, the diffusion, and the pH value based on the theories of the diffusion reaction of thermodynamics and the chemotactic reaction of biology in this paper. The stability of the model is studied by analyzing the growth rate and the frequency of the solutions of a small perturbation. When different parameter values are set, the effect of every parameter on the stability of the system is analyzed in detail. With a number of data analysis, the conclusion that the stability is mainly affected by two parameters, namely, the initial density of bacteria and the consumption rate of foods by bacteria when they are given the large enough values.**

***Index Terms*—microbial growth, dynamic model, stability, chemotaxis, organism**

I. Introduction

Biofouling and bio-corrosion has become a serious problem in the industrial cooling water system. Microbial colonization of industrial equipments results from the formation of biofilms which are made of bacteria, extracellular polymeric substances (EPS) and mainly water. When the environment is suitable, the microorganisms are able to grow and reproduce rapidly. Once these microorganisms are excessively deposited, they will ultimately form the biofouling and attached to the surface of the pipelines and industrial equipments. This will increase thermal resistance, reduce the heat exchange capacity of heat exchanger, accelerate the corrosion of structural mental and alloys and reduce the production efficiency. Consequently, it is absolutely necessary to study and predict the microbial growth in power plant and the industrial cooling water system.

Microorganisms usually live in the complicated environment. The interaction of microorganisms can lead to long-term dynamical systems. A number of mathematical models about microbial growth based on dynamics have be developed. Keller & Segel developed the K-S model and analyzed the onset condition of instability [1, 2] based on the chemotaxis behavior of ameba. Subsequently, a variety of mathematical model is developed based on the K-S model. Men et al. [3] developed a dynamic model of microbial growth by considering the bacterial chemotaxis, diffusion and pH value and made a simple analysis on the stability of the model. Li et al. [4] made further analysis on the stability of the model in detail by ignoring pH value, ignoring the chemotaxis and the whole system respectively.

The present paper provides another perspective to insight into the stability of the model. The effect of every parameter in the governing equations on the stability on the system is studied by a number data experiments. Finally, some significant conclusions are drawn that the initial density of bacteria and the consumption rate of foods by bacteria are two main influencing factors when their values reach enough large.

In this paper, the content is organized as follows: The mathematical model is presented in Section 2. The small perturbations analysis is made in Section 3. The stability analysis of the system is given in Section 4 and concluding remarks are addressed in Section 5.

II. Mathematical Model

Lin built continuous partial differential equations about the amoeba’s aggregation and diffusion behavior [5]. He accounted for the relationship between the density of bacteria and the concentration of attractant. But so far, the pH value has not been considered as an influential factor on the system in the control equations. The following will be a detailed description of the relationship between the density of bacteria, the concentration of food (also called the concentration of attractant) and the pH value, and the control equations of the model are governed by

$$\frac{\partial u}{\partial t} = \nabla\cdot(\mu\nabla u) - \nabla\cdot(\kappa u\nabla f) + auf - bu + \psi(\varphi)u \tag{1}$$

$$\frac{\partial f}{\partial t} = -su + gu - hf + \nabla\cdot(\nu\nabla f) \tag{2}$$

$$\frac{\partial \varphi}{\partial t} = \nabla\cdot(\eta\nabla\varphi) + \varphi_s \tag{3}$$

Where u is the density of bacteria, f is the concentration of food or the concentration of nutrition, here it is defined as the concentration of attractant; φ is the pH value of the aqueous solution where bacteria lives; φ_s is the source of the pH value and t is the time.

Equation (1) denotes the increase rate of bacterial density, where the first term on the right represents the diffusion reaction which always disperses along the

doi:10.4304/jcp.5.5.741-748

gradient in the opposite direction of bacterial density and it represents the motion of the bacteria in the absence of attractant. The motility parameter, μ , represents the acuteness degree of the random motion of bacteria, here called the bacterial diffusion coefficient. The second term denotes the chemotactic reaction of bacteria, describes the behavior that bacteria moves towards any element which is conducive for their survival, here represented as f . κ is a measure of the strength of chemotaxis, and is termed as the chemotactic coefficient. The third term describes the reproduction of bacteria which is influenced by two factors, one is maternal bacteria and the other is nourishment, namely termed attractant, where a is reproduction rate per bacteria. The fourth term describes the bacterial natural death which relates to the number of bacterial population, b denotes the mortality rate per bacteria. The fifth term describes the contribution of the pH value to the microbial growth, and it is represented as a function $\psi(\varphi)$ which can be taken as Gauss function or rectangular function [6].

Equation (2) denotes the secretion rate of attractant, where the first term describes the consumption of attractant by bacteria, s denotes the consumption rate of the foods by bacteria; the second term describes that bacterial excretion and the dead bacterial individual become into foods, g denotes the conversion rate; the third term describes attractant degradation, h denotes the degradation rate of attractant. The fourth term describes the diffusion response of the attractant, ν denotes the diffusion coefficient of the attractant (also could be other factors which are conducive for bacterial growth). Equation (3) represents the increase rate of the pH value, where the first term describes the diffusion reaction of the pH value, η denotes the diffusion coefficient of the pH value. The second term denotes the source of the pH value. Since bacteria lives in the thin layer of biofilm, we believe that the influence of convection is far less than the diffusion reaction [7], thus the influence of the diffusion are ignored in (3). In addition, the parameters such as μ , κ , a , b , s , g , h , ν , η can also be constants, or could be functions depending on the parameters, u , f , φ .

III. The Analysis of Small Perturbation

The constants u_0 , f_0 and φ_0 are assumed as a solution for (1)~(3), then they will meet the following formulas

$$af_0 - b + \psi(\varphi_0) = 0 \tag{4}$$

$$-su_0 + gu_0 - hf_0 = 0 \tag{5}$$

$$\varphi_s(u_0, \varphi_0, f_0) = 0 \tag{6}$$

At that time, the reproduction rate of bacteria and the positive contribution of the pH value to the bacterial growth are equal to the death rate of bacteria. Moreover, the secretion rate or the generation rate of attractant can be offset by the consumption rate of attractant by bacteria. Therefore, the system is in a uniform state. That is, the system is in an equilibrium point. In the following, we will make a study on the stability of the system when it is disturbed by a small perturbation and deviates from the equilibrium point. If the small perturbation tends to disappear, the system is stable; if the perturbation is enlarged and becomes very powerful, which makes the system deviate from the equilibrium point into the collapse or entering another equilibrium point, the system is unstable.

Supposing at a certain point, a small perturbation appears, $u^{'}$, $f^{'}$ and $\varphi^{'}$ respectively describes the perturbation of bacterial density, the perturbation of attractant concentration and the perturbation of the pH value which deviate from the uniform state of the system, and they are all small quantities. Here the disturbed physical quantities are termed as $u_0 + u^{'}$, $f_0 + f^{'}$, $\varphi_0 + \varphi^{'}$ respectively. We put them into the (1) ~ (3), then subtract the basic field and ignore the non-linear quantities resulting from the product of two small quantities, consequently obtaining

$$\frac{\partial u^{'}}{\partial t} = \nabla \cdot (\mu \nabla u^{'}) - \nabla \cdot (\kappa u_0 \nabla f^{'}) + u_0[af + \frac{\partial \psi}{\partial \varphi}(\varphi_0)\varphi^{'}] \tag{7}$$

$$\frac{\partial f^{'}}{\partial t} = -su^{'} + gu^{'} - hf^{'} + \nabla \cdot (\nu \nabla f^{'}) \tag{8}$$

$$\frac{\partial \varphi^{'}}{\partial t} = \nabla \cdot (\eta \nabla \varphi^{'}) + \frac{\partial \varphi_s}{\partial u} u^{'} + \frac{\partial \varphi_s}{\partial f} f^{'} + \frac{\partial \varphi_s}{\partial \varphi} \varphi^{'} \tag{9}$$

For the above-mentioned small perturbation linear partial differential equations, the linear stability is analyzed with the normal module method [8]. Assuming the solutions for (7) ~ (9) are in the forms by the following

$$u^{'} = C_1 \sin(qx) e^{\sigma t} \tag{10}$$

$$f^{'} = C_2 \sin(qx) e^{\sigma t} \tag{11}$$

$$\varphi^{'} = C_3 \sin(qx) e^{\sigma t} \tag{12}$$

Here C_1 , C_2 , C_3 is constant respectively. The assumed solutions of the above are put into the small perturbation equations, obtaining

$$\sigma C_1 = -\mu q^2 C_1 + \kappa u_0 q^2 C_2 + u_0 a C_2 + u_0 \psi^{'}(\varphi_0) C_3 \tag{13}$$

$$\sigma C_2 = (g - s) C_1 - h C_2 - \nu q^2 C_2 \tag{14}$$

$$\sigma C_3 = -\eta q^2 C_3 + \frac{\partial \varphi_s}{\partial u} C_1 + \frac{\partial \varphi_s}{\partial u} C_2 \tag{15}$$

These equations are the homogeneous equations of C_1 , C_2 , C_3 ,and the conditions for the solvability of (13)~(15) are

$$\begin{vmatrix} \sigma+\mu q^2 & -\kappa u_0 q^2 - u_0 a & -\mu\psi^{'}(\varphi_0) \\ -(g-s) & \sigma+h+\nu q^2 & 0 \\ -\frac{\partial\varphi_s}{\partial u}(u_0,\varphi_0,f_0) & -\frac{\partial\varphi_s}{\partial f}(u_0,\varphi_0,f_0) & \sigma+\eta q^2-\frac{\partial\varphi_s}{\partial\varphi}(u_0,\varphi_0,f_0) \end{vmatrix} = 0 \qquad (16)$$

Expand (16), we'll get the following cubic equation,

$$\sigma^3 + A\sigma^2 + B\sigma + C = 0 \qquad (17)$$

That is the general spectrum relationship. Here

$$A = (\mu+\nu+\eta)q^2 + h - \frac{\partial\varphi_s}{\partial\varphi} \qquad (18)$$

$$B = (\mu\nu+\mu\eta+\nu\eta)q^4 + [\mu h + h\eta - \kappa u_0(g-s-\mu\frac{\partial\varphi_s}{\partial\varphi} -\nu\frac{\partial\varphi_s}{\partial\varphi})q^2 - h\frac{\partial\varphi_s}{\partial\varphi}] - \mu\psi^{'}(\varphi_0)\frac{\partial\varphi_s}{\partial u} - (g-s)u_0 a \qquad (19)$$

$$C = \mu\nu\eta q^6 + (\mu h\eta - \kappa u_0\eta g + \kappa u_0\eta s - \mu\nu\frac{\partial\varphi_s}{\partial\varphi})q^4 + [(\kappa u_0 g -\kappa u_0 s - \mu h)\frac{\partial\varphi_s}{\partial\varphi} - \mu\nu\psi^{'}\frac{\partial\varphi_s}{\partial u} - \eta u_0 a(g-s)]q^2 -(g-s)\mu\psi^{'}\frac{\partial\varphi_s}{\partial f} - \mu h\psi^{'}\frac{\partial\varphi_s}{\partial u} + u_0 a(g-s)\frac{\partial\varphi_s}{\partial\varphi} \qquad (20)$$

The general solutions for (17) are

$$\sigma_1 = -\frac{A}{3} - \frac{2^{1/3}(-A^2+3B)}{3\chi} + \frac{\chi}{3\times 2^{1/3}} \qquad (21)$$

$$\sigma_2 = -\frac{A}{3} + \frac{(1+\sqrt{3}i)(-A^2+3B)}{3\times 2^{2/3}\chi} - \frac{(1-\sqrt{3}i)\chi}{6\times 2^{1/3}} \qquad (22)$$

$$\sigma_3 = -\frac{A}{3} + \frac{(1-\sqrt{3}i)(-A^2+3B)}{3\times 2^{2/3}\chi} - \frac{(1+\sqrt{3}i)\chi}{6\times 2^{1/3}} \qquad (23)$$

here

$$\chi = [-2A^3 + 9AB - 27C + 3\sqrt{3}(-A^2B^2 + 4B^3 + 4A^3C -18ABC + 27C^2)^{1/2}]^{1/3} \qquad (24)$$

It is obvious that the solutions seem very simple in the forms, but it is very complicated to look for neutral curves. In order to simplify the problem without loss of generality, some of the parameters can be given up as the time goes by, that is an assumption that some parameters have been known or they are not important, then the influence of the factors on the perturbation can be further studied.

Ⅳ. The Stability Analysis of the System

In general, when the small perturbation problem is researched, we only concern the development of state of the substance concentration after equilibrium state is disturbed for a short time, not involve the evolution of a long time. In other words, we only concern the transient development trend of the perturbation. Hence, assumptions should be taken that there is no the reproduction or death of bacterial individuals and there is no dead bacteria being transformed into attractant or there is no attractant's degradation in this course, so the parameters a, b, g, h and the function ψ can all be taken to be zero. In addition, φ_s is a complex physical quantity, so it is temporarily assumed as a constant in order to simplify the problem. Accordingly, Equation (16) is transformed into

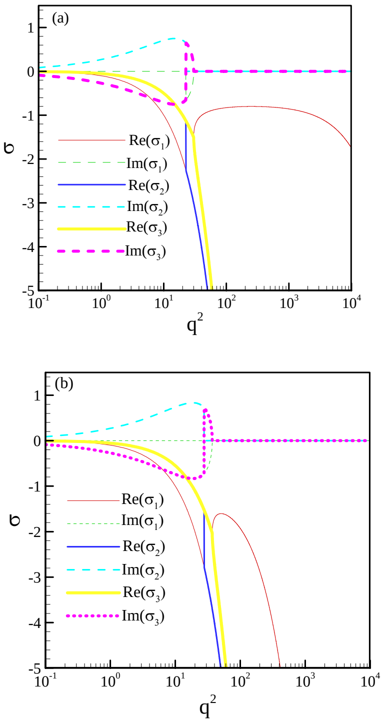
$$\sigma^3 + (\nu+\mu+\eta)q^2\sigma^2 + [(\mu\nu+\mu\eta+\nu\eta)q^4 + s\kappa u_0 q^2]\sigma + \nu\mu\eta q^6 + \eta s\kappa u_0 q^4 = 0 . \qquad (25)$$

From (25), we can see that the solutions are affected by seven parameters (μ, ν, η, κ, s, u_0, q^2). Therefore, it is significant to study the effects of these parameters on the stability of perturbations.

A The Effect of Bacterial Diffusion Rate on the System's Stability

(1) While $u_0 = 0.1\times10^{-4}/cm^2$, $s = 0.1\times10^{-4}cm^2/s$

The effects of $\mu = 0.0001\times10^{-4}$, 0.01×10^{-4}, 1.0×10^{-4}, 10.0×10^{-4}, $10000.0\times10^{-4}cm^2/s$ on the growth rate and the frequency of the perturbations when $\kappa = 7.5\times10^{-4}cm^2/s$, $\nu = 0.1\times10^{-4}cm^2/s$ and $\eta = 0.1\times10^{-4}cm^2/s$ are shown in Figure 1. (a)—(d).



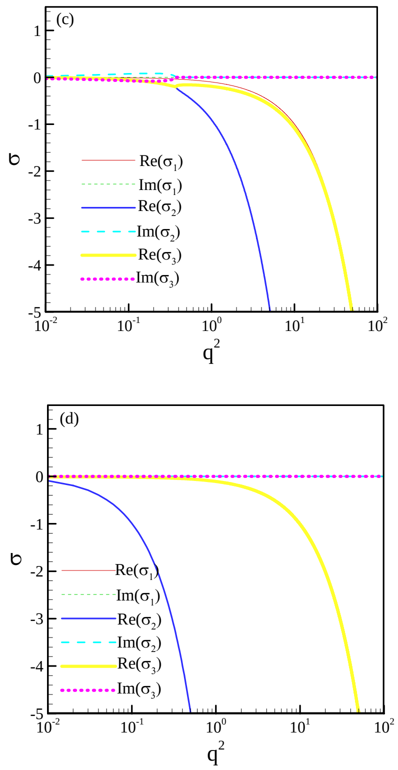
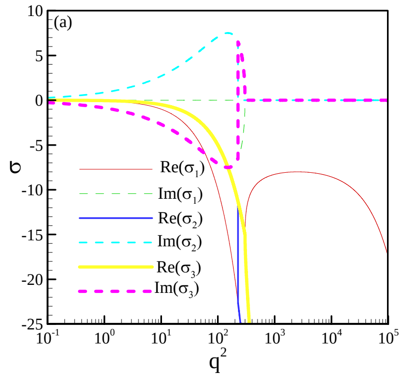


Figure 1. The growth rate and the frequency of the solutions when $\kappa = 7.5\times10^{-4}cm^2/s$, $\nu = 0.1\times10^{-4}cm^2/s$ and $\eta = 0.1\times10^{-4}cm^2/s$ for $u_0 = 0.1\times10^{-4}/cm^2$, $s = 0.1\times10^{-4}cm^2/s$. (a) $\mu = 0.0001\times10^{-4}cm^2/s$; (b) $\mu = 0.01\times10^{-4}cm^2/s$; (c) $\mu = 1.0\times10^{-4}cm^2/s$; (d) $\mu = 10.0\times10^{-4}cm^2/s$; (f) $\mu = 10000.0\times10^{-4}cm^2/s$

. Figure 1 shows the effects of different μ on the growth rate and the frequency of the perturbations. It can be seen that the system is always stable when the initial density of the bacteria and the consumption rate of foods by bacteria is very low, and it is influenced very little by microbial diffusion under this circumstance. But the frequency of the system is largely influenced by that.

(2) While $u_0 = 1.0\times10^{-4}/cm^2$, $s = 0.1\times10^{-4}cm^2/s$

The effects of $\mu = 0.0001\times10^{-4}$, $1.0\times10^{-4}cm^2/s$ on the growth rate and the frequency of the perturbations when $\kappa = 7.5\times10^{-4}cm^2/s$, $\nu = 0.1\times10^{-4}cm^2/s$ and $\eta = 0.1\times10^{-4}cm^2/s$ are shown in Figure 2. (a)—(b).



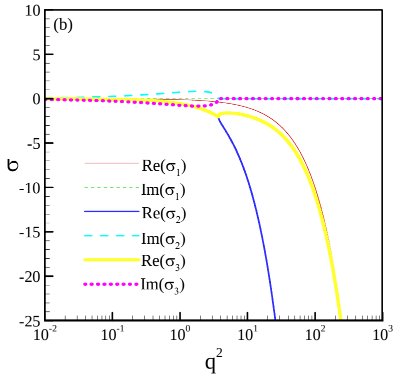


Figure 2. The growth rate and the frequency of the solutions when $\kappa = 7.5\times10^{-4}cm^2/s$, $\nu = 0.1\times10^{-4}cm^2/s$ and $\eta = 0.1\times10^{-4}cm^2/s$ for $u_0 = 0.1\times10^{-4}/cm^2$, $s = 0.1\times10^{-4}cm^2/s$. (a) $\mu = 0.0001\times10^{-4}cm^2/s$; (b) $\mu = 1.0\times10^{-4}cm^2/s$.

The system is still stable shown in Figure 2. But the frequency is larger in Figure 2(a) than that in Figure 1(a) when the initial density of bacteria (u_0) increases with the other parameters not varying. This interprets that the initial density of bacteria influences the frequency of the perturbations, which makes the system is stable with oscillation.

(3). While $s = 0.1\times10^{2}cm^2/s$, $u_0 = 0.1\times10^{4}/cm^2$

The effects of $\mu = 0.0001\times10^{-4}$, $1000.0\times10^{-4}cm^2/s$ on the growth rate and the frequency of the perturbations when $\kappa = 7.5\times10^{-4}cm^2/s$, $\nu = 0.1\times10^{-4}cm^2/s$ and $\eta = 0.1\times10^{-4}cm^2/s$ are shown in Figure 3. (a)—(b).

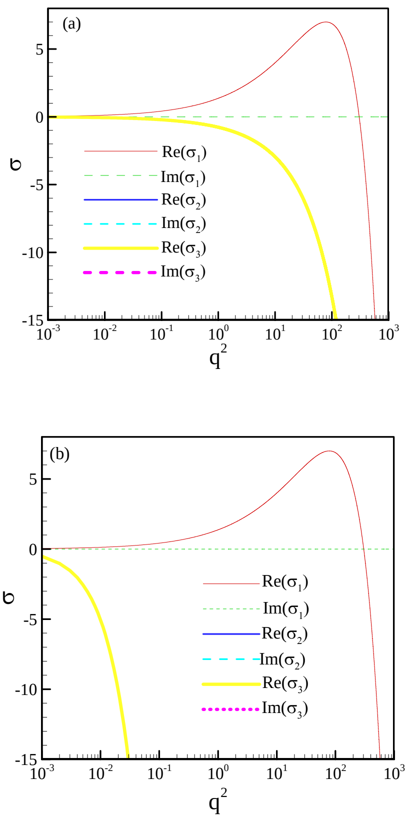


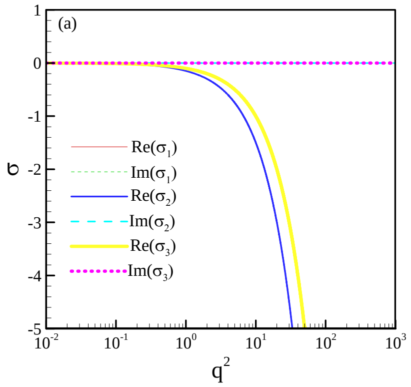
Figure 3. The growth rate and the frequency of the solutions when $\kappa = 7.5\times10^{-4} cm^2/s$, $\nu = 0.1\times10^{-4} cm^2/s$ and $\eta = 0.1\times10^{-4} cm^2/s$ for $s = 0.1\times10^{2} cm^2/s$, $u_0 = 0.1\times10^{4}/cm^2$. (a) $\mu = 0.0001\times10^{-4} cm^2/s$; (b) $\mu = 1.0\times10^{-4} cm^2/s$.

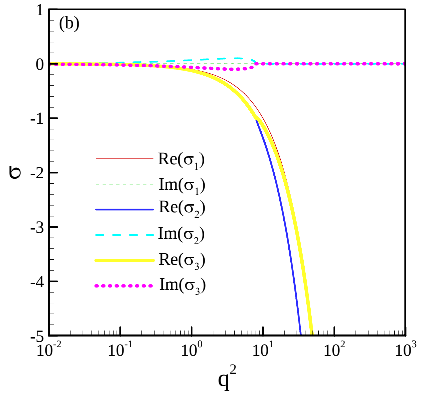
From Figure 3, we can see that the unstable region of the system occurs when the parameter s increases in Figure 3 compared with that in Figure 2. The unstable region extends with the increase of μ under. This phenomenon interprets that the more intense microbial diffusion becomes, the more unstable the system becomes.

B. The Effect of Bacterial Chemotaxis on the System's Stability

(1). While $u_0 = 0.1\times10^{-4}/cm^2$, $s = 0.1\times10^{-4} cm^2/s$

The effects of $\kappa = 0.005\times10^{-4}$, $\kappa = 0.5\times10^{-4}$ $\kappa = 5.0\times10^{-4} cm^2/s$ on the growth rate and the frequency of the perturbations when $\mu = 0.15\times10^{-4} cm^2/s$, $\nu = 0.1\times10^{-4} cm^2/s$ and $\eta = 0.1\times10^{-4} cm^2/s$ are shown in Figure 4. (a)—(c).





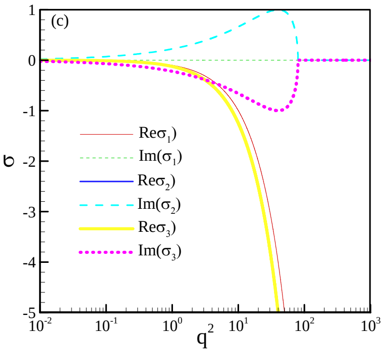


Figure 4. The growth rate and the frequency of the solutions when $\mu = 0.15\times10^{-4} cm^2/s$, $\nu = 0.1\times10^{-4} cm^2/s$ and $\eta = 0.1\times10^{-4} cm^2/s$ for $u_0 = 0.1\times10^{-4}/cm^2$, $s = 0.1\times10^{-4} cm^2/s$. (a) $\kappa = 0.005\times10^{-4} cm^2/s$; (b) $\kappa = 0.5\times10^{-4}/s$; (c) $\kappa = 5.0\times10^{-4} cm^2/s$.

Figure 4 shows that the system is always stable with the increase of κ when the initial density of bacteria and the consumption of foods by bacteria is low. Under this circumstance, the chemotaxis affects little on the stability of the system, but affects large on the frequency.

(2).While $u_0 = 1000.0\times10^{-4}/cm^2$, $s = 0.1\times10^{-4} cm^2/s$

The effects of $\kappa = 0.005\times10^{-4}$, $\kappa = 0.05\times10^{-4} cm^{-2}/s$ on the growth rate and the frequency of the perturbations when $\mu = 0.15\times10^{-4} cm^2/s$, $\nu = 0.1\times10^{-4} cm^2/s$ and $\eta = 0.1\times10^{-4} cm^2/s$ are shown in Figure 5 (a)—(b).

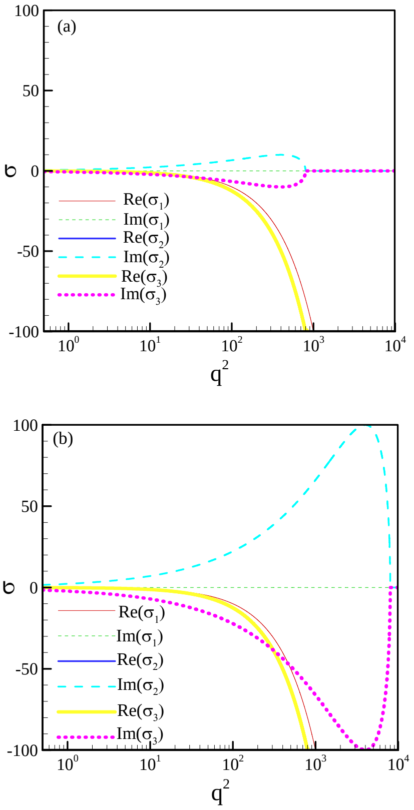


Figure 5. The growth rate and the frequency of the solutions when $\mu = 0.15\times10^{-4} cm^2/s$, $\nu = 0.1\times10^{-4} cm^2/s$ and $\eta = 0.1\times10^{-4} cm^2/s$

for $u_0 = 1000.0\times10^{-4}/cm^2$, $s = 0.1\times10^{-4}cm^2/s$. (a) $\kappa = 0.005\times10^{-4}cm^2/s$; (b) $\kappa = 0.05\times10^{-4}cm^{-2}/s$.

The effect of the bacterial chemotaxis on the stability is similar with that in Figure 5. Here the initial density of bacteria is much larger than that in Figure 5, so the increase of u_0 only affects the frequency of the system when the other parameters do not vary.

(3). While $u_0 = 0.1\times10^{-4}/cm^2$, $s = 1000.0\times10^{-4}/cm^2$

The effects of $\kappa = 0.005\times10^{-4}cm^2/s$, $\kappa = 0.05\times10^{-4}cm^{-2}/s$ on the growth rate and the frequency of the perturbations when $\mu = 0.15\times10^{-4}cm^2/s$, $\nu = 0.1\times10^{-4}cm^2/s$, $\eta = 0.1\times10^{-4}cm^2/s$, $u_0 = 0.1\times10^{-4}/cm^2$ are shown in Figure 6. (a)—(b).

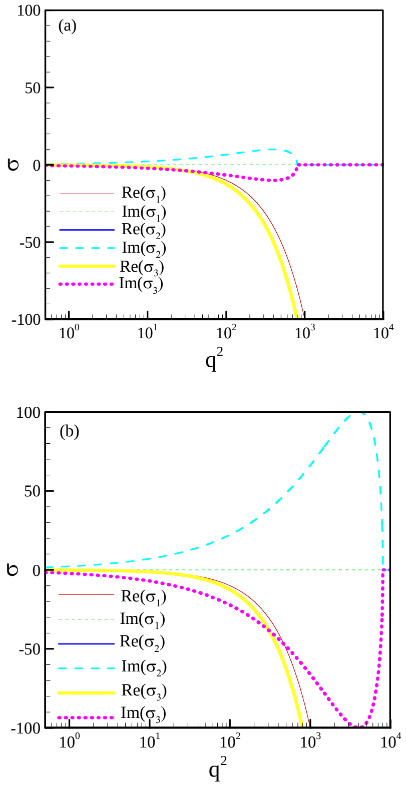


Figure 6. The growth rate and the frequency of the solutions when $\mu = 0.15\times10^{-4}cm^2/s$, $\nu = 0.1\times10^{-4}cm^2/s$ and $\eta = 0.1\times10^{-4}cm^2/s$ for $u_0 = 0.1\times10^{-4}/cm^2$, $s = 1000.0\times10^{-4}cm^2/s$. (a) $\kappa = 0.005\times10^{-4}cm^2/s$; (b) $\kappa = 0.05\times10^{-4}cm^{-2}/s$.

The same conclusion is drawn in Figure 6 that the increase of bacterial chemotaxis (κ) and the consumption rate of foods by bacteria (s) do not affect the stability, they only affect the frequency of the perturbations.

C. The Effect of Diffusin Rate of the pH Value on the System's Stability

(1). While $u_0 = 0.1\times10^{-4}/cm^2$, $s = 0.1\times10^{-4}cm^2/s$

The effects of η=$0.01\times10^{-4}cm^2/s$, η=$10.0\times10^{-4}cm^2/s$ on the growth rate and the frequency of the perturbations when $\mu = 0.15\times10^{-4}cm^2/s$, $\nu = 0.1\times10^{-4}cm^2/s$, $\kappa = 7.5\times10^{-4}cm^2/s$, are shown in Figure 7. (a)—(b).

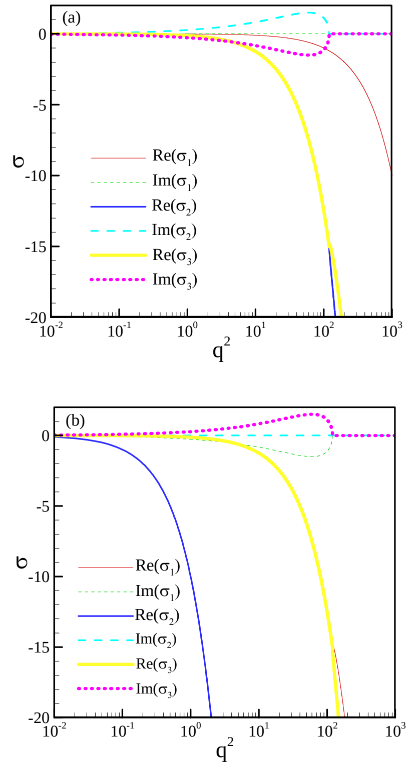
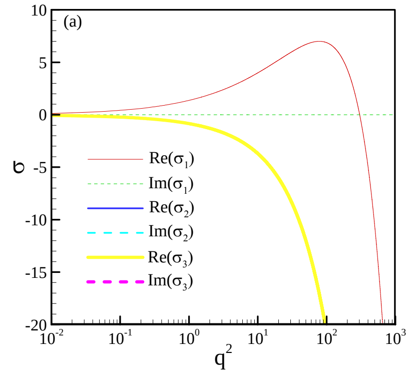


Figure 7. The growth rate and the frequency of the solutions when $\mu = 0.15\times10^{-4}cm^2/s$, $\nu = 0.1\times10^{-4}cm^2/s$ and $\kappa = 7.5\times10^{-4}cm^2/s$ for $u_0 = 0.1\times10^{-4}/cm^2$, $s = 0.1\times10^{-4}cm^2/s$. (a) η=$0.01\times10^{-4}cm^2/s$; (b) η=$10.0\times10^{-4}cm^2/s$.

Figure 7 shows that the system is always when η increases. It interprets the diffusion of the pH value has little effect on the stability of the system when the initial density of bacteria and the consumption rate of foods by bacteria is very low.

(2). While $u_0 = 0.1\times10^{4}/cm^2$, $s = 0.1\times10^{2}cm^2/s$

The effects of η=$0.1\times10^{-4}cm^2/s$, η=$1.0\times10^{-4}cm^2/s$, η=$10.0\times10^{-4}cm^2/s$ on the growth rate and the frequency of the perturbations, when $\mu = 0.15\times10^{-4}cm^2/s$, $\nu = 0.1\times10^{-4}cm^2/s$, $\kappa = 7.5\times10^{-4}cm^2/s$, are shown in Figure 8 (a)—(c).



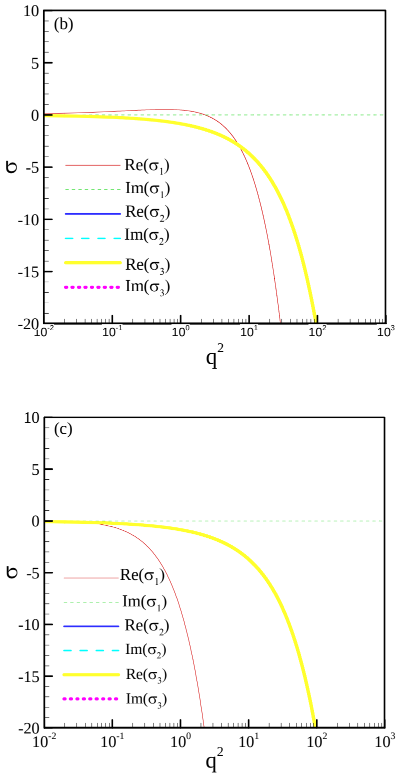
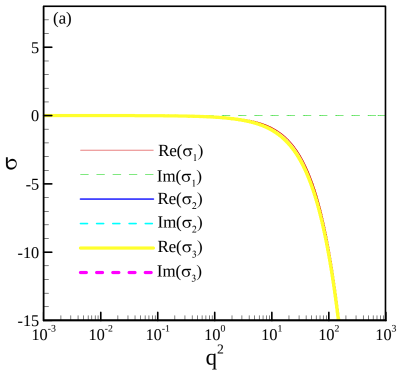


Figure 8. The growth rate and the frequency of the solutions when $\mu = 0.15\times10^{-4} cm^2/s$, $\nu = 0.1\times10^{-4} cm^2/s$ and $\kappa = 7.5\times10^{-4} cm^2/s$ for $u_0 = 0.1\times10^4/cm^2$, $s = 0.1\times10^2 cm^2/s$. (a) η=$0.1\times10^{-4} cm^2/s$; (b) η=$1.0\times10^{-4} cm^2/s$; (c) η=$10.0\times10^{-4} cm^2/s$.

Figure 8 illustrates the stability of the system is affected by η when the initial density of bacteria and the consumption rate of foods by bacteria is very high. Under this circumstance, the unstable region shrinks with the increase the intensity of the diffusion of pH value.

D The Effect of the Consumption Rate of Foods by Bacteria on the System's Stability

The effects of s=100.0×10^{-4} , $10000.0\times10^{-4}/cm^2$, 100000.0×10^{-4} on the growth rate and the frequency of the perturbations when $\mu = 0.1\times10^{-4} cm^2/s$, $\nu = 0.1\times10^{-4} cm^2/s$, $\kappa = 7.5\times10^{-4} cm^2/s$, η=$0.1\times10^{-4} cm^2/s$, $u_0 = 0.1\times10^4/cm^2$ are shown in Figure 9 (a)—(c).



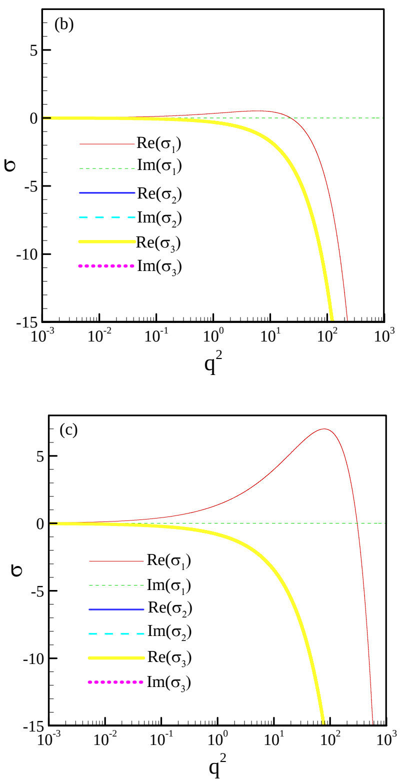
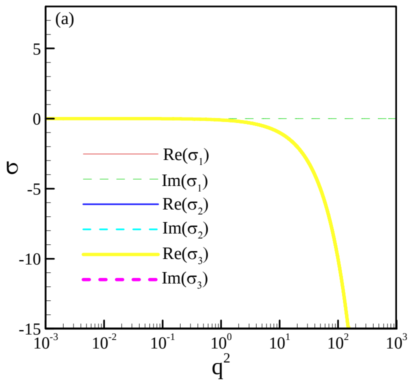


Figure 9. The growth rate and the frequency of the solutions when $\mu = 0.1\times10^{-4} cm^2/s$, $\nu = 0.1\times10^{-4} cm^2/s$, $\kappa = 7.5\times10^{-4} cm^2/s$, η=$0.1\times10^{-4} cm^2/s$, $u_0 = 0.1\times10^4/cm^2$. (a) s=$100.0\times10^{-4}/cm^2$; (b) s=$10000.0\times10^{-4}/cm^2$; (c) s=$100000.0\times10^{-4} cm^2$.

Figure 9 shows that the stability of the system is largerly affected by the consumption rate of foods by bacteria when the initial density of bacteria is very high. The system becomes unstable with the increase of s .

E The Effect of the Initial Density of Bacteria on the System's Stability

The effects of $u_0 = 1.0\times10^{-2}$, 1.0×10^2 , $1.0\times10^3/cm^2$ when $\mu = 0.1\times10^{-4} cm^2/s$, $\nu = 0.1\times10^{-4} cm^2/s$, $\kappa = 7.5\times10^{-4} cm^2/s$, η=$0.1\times10^{-4} cm^2/s$, $s = 0.1\times10^2 cm^2/s$ on the growth rate and the frequency of the perturbations are shown in Figure 10 (a)—(c).



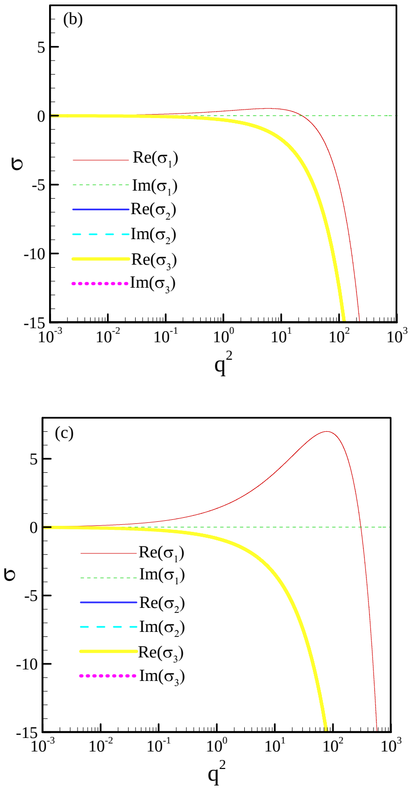


Figure 10. The growth rate and the frequency of the solutions when $\mu = 0.1\times10^{-4} cm^2/s$, $\nu = 0.1\times10^{-4} cm^2/s$, $\kappa = 7.5\times10^{-4} cm^2/s$, η=$0.1\times10^{-4} cm^2/s$, $s = 0.1\times10^{2} cm^2/s$. (a) $u_0 = 1.0\times10^{-2}/cm^2$; (b) $u_0 = 1.0\times10^{2}/cm^2$; (c) $u_0 = 1.0\times10^{3}/cm^2$.

The same conclusion is drawn in Figure 10 as in Figure 9. The initial density of bacteria also largely affects the stability of the system. When u_0 increases, the unstable region extends.

V CONCLUSIONS

A nonlinear dynamic model of microbial growth is developed and the stability of small perturbation equations are analyzed under the control of seven variable parameters μ , ν , η , s , κ , u_0 , q^2 in the present paper. The effects of every parameter on the growth rate and the frequency of the solutions are studied when different values are set to them. By a large number of data being analyzed, some conclusions have be drawn as follows: (1) when the initial density of bacteria and the consumption rate of foods by bacteria is very low, the system is always stable and it is affected very little by whichever parameter such as microbial diffusion rate μ , the bacterial chemotaxis κ or the diffusion rate of the pH value η . (2) When the initial density of bacteria and the consumptions rate of foods by bacteria is enough high, the stability of the system is affected largely by microbial diffusion rate μ , the bacterial chemotaxis κ and the diffusion rate of the pH value η . When μ , κ , or η increases, the stability becomes unstable gradually. (3) Both the initial density of bacteria u_0 and the consumptions rate of foods by bacteria have a large influence on the stability of the system. When u_0 is very large, the stability of the system becomes unstable with the increase of s . Similarly, when s is very large, the stability of the system becomes unstable with the increase of u_0 .

ACKNOWLEDGMENT

This work is supported both by CNSF # 60604023 and by National Basic Research Program of China (973 Program) #2007CB206904 to Hong Men.

REFERENCES

[1] E.F. Keller, and L.A. Segel, "Initiation of slime mold aggregation viewed as an instability," *J. theor. Biol.*, vol. 26, pp.399–415, March 1970.

[2] E.F. Keller, and L.A. Segel, "Model for chemotaxis," *J. theor. Biol.* , vol. 30, pp. 225–234, February 1971.

[3] H. Men, Y. H. Li, and F. B. Tian, "Microbial growth and stability analysis," *ECS2009*, Wuhan, March 2009.

[4] Y. H. Li, F. B. Tian, J. Wu, and H. Men, "The nonlinear microbial growth model and stability anslysis," *J. Theor. Biol.* (submitted)

[5] C. C. Lin, and L. A. Segel, *Mathematics Applied to Deterministic Problems in the Natural Sciences*, Macmillan, New York, 1974.

[6] Y. H. Tan, Z. X. Wang, and K. C. Marshall, "Modeling pH effects on microbial growth: a statistical thermodynamic approach," *Biotechnol. Bioeng.*, vol. 59, pp. 724-731, 1998.

[7] D. Beer, and P. Stoodley, "Relation between the structure of an aerobic biofilm and transport phenomena," *Wat. Sci. Tech.*, vol 32, pp.11-18, 1995.

[8] W. Criminale, T.L. Jacksen, and R.D. Joslin, *Theory and Computation of Hydrodynamic Stability*, Cambridge Univ.Press, 2003.

Hong Men was born in Jilin, China, in 1973. He received a Bachelor of Electronics and Informational System from Northeast Normal University, Changchun, China, in 1996, a Master of Electric Power System and Automation from Northeast Institute of Electric Power Engineering, Jilin, China, in 2002, and a Doctor of Biomedical Engineering from Zhejiang University, Hangzhou, China, in 2005. Now he is an associate professor in Northeast Dianli University. His current research areas are biofilm and biofouling, biosensor and pattern recognition technique.

Yuhong Li was born in Qingdao, China, in 1972. She graduated in Electronic Technology & Computer Application from University of Science and Technology of China, Hefei, China, in 1997, and presently is a postgraduate student in Northeast Dianli University, Jilin, China. Her main interests are dynamic model, microbial growth, nonlinear system and automaton.

Fangbao Tian was born in Yunnan, China, in 1984. He graduated in Theoretical and Applied Mechanics at the Department of Modern Mechanics, University of Science and Technology of China, Hefei, China, in 2006, and presently is a PhD Candidate in University of Science and Technology of China. His main interests are computational fluid dynamics, bio-fluid mechanics, turbulence research, non-Newtonian fluids and flow stability.

Adding Branching Temporal Dimension to Qualitative Coalitional Games with Preference

Cheng Bailaing[1,2] Zeng Guosun[1]
[1]Department of computer Science and Technolog, Tongji University, Shanghai, China
Email: cblcbl2002@hotmail.com, gszeng@mail.tongji.edu.cn,

Jie Anquan[2]
[2]College of computer information and engineering, Jiangxi normal University, Nanchang, China
Email: jaq@jxnu.edu.cn

***Abstract*—Qualitative Coalitional Games (QCGs), as a variation of coalitional games, is to investigate agents' strategies and behaviors in cooperating games. Each agent has a set of goals as its desires and will be satisfied if at least one of its desires is achieved by executing some strategies in a coalition, otherwise be unsatisfied. After introducing QCGs, we add preference to QCGs framework to enable that every agent has the ability to join the best coalition for achieving its preferences goals (QCGPs). In order to make a formal description and reason about repeated coalitional games, the paper will study Branching Temporal Qualitative Coalitional Games with preferences. Computational Tree Logic (CTL) is used for QCGPs with complete axiomatisation of it, denoted as CTQCGPs. Further more, this paper analyses the expression power, complexity, and some characteristic of CTQCGPs.**

***Index Terms*—Coalition games, Multi-agents Systems, Modal logic, Artificial Intelligence.**

I. Introduction

The study of social software is a hot topic among research communities, including computer scientists, game theorists and philosophers. The key idea of it is regarding the social process as a computer program, and then using formal methods to analyze, design and validates social procedures just the same way in computer programs [1]. On the other hand, the multi-agents systems have gained the attention of game theory and artificial intelligence. So using formal methods, such as model and logic are an effective way for games which are often studied by mathematics tools. With the idea, coalitional Games (CGs) [2.3] are regarded as a natural tool for modeling goal-oriented multi-agents systems. When the goals of agents can be achieved with transferable payoffs of which the expression of payoff is not a numeric value, we have to use qualitative ways to solve the problem, Wooldridge.et.al[4] first introduces QCGs and pays an attention to the computation complexity of QCGs for decision problem and make the definition of the satisfaction of agents. In QCGs, each agent has a set of goals as its desires and will be satisfied in a coalition by performing some strategies. The means of an agent's satisfaction is that a coalition including that agent achieves a set of goals which at least include one of the elements of that agent's goals. [5, 6]Add preference to QCGs without temporal concept, what the means of preference of agent is that in any two goals of the agent, it prefers one to the other goal. In order to systematically study the multi-agents systems [7], repeated games [8] are an important way. Line Temporal logic(LTL) is used for repeated games in QCGs(TQCGs) by iterating games in time serial[9], but the expression power of LTL is only useful in universal quantifier. This paper will use CTL for QCGPs(CTQCGPs) for existential and universal quantifier.

This paper is organized as follows: Section Ⅱ introduces QCGPs with a formal description, section Ⅲ defines logic for expressing the properties of individual QCGPs, and the language for QCGPs is also given. In section Ⅳ we built CTQCGPs. Sone chatacteristic of CTQCGPs is given in Ⅴ. Conclusion and future work will be given in section Ⅵ.

II. Qualitative Coalitional Games with Preferences

Coalitional games were introduced in [4] for interpreting cooperative interactions in games, just like "which coalition I should join in". Although coalitional games are a very good model for numeric value in which every agent's payoff can be expressed as a numeric value, unfortunately, it can not enable the payoff of a goals set to be expressed in that way. The new situation assumes that every agent has a set of goals as its desires and achieves its desires by attending a coalition to get some goals. To a different agent, in two goals, it prefers a goal to the other between two goals. The abstract description of coalitional games is given in [8], the formal model of QCGs is first presented in [4], and QCGPs is introduced in [5]. These models pay attention to how the coalition can be formal without caring for the concrete strategy and how to compute the payoff of agents.

QCGPs include a non-empty, finite set $A=(1, \dots m)$ of agents, each agent has a set of goals $G_i \subseteq G$. Here G is a goals set of all agents, and the elements of G_i has a partial order relation for interpreting the preference of agent i.

Definition 1 Qualitative Coalitional Games with Preferences (QCGPs) is a $2n+3$ tuple

doi:10.4304/jcp.5.5.749-756

$$\Gamma=<A,\ G\ ,G_1,\ldots,G_n,\ V,\ \rhd_1,\ldots,\rhd_n>$$

where. A is a finite, non-empty set of agents; $G=\{g_1,\ldots,g_m\}$is a set of possible goals; $G_i \subseteq G$ is the set of goals for agent $i \in A$; $V: 2^{A} \rightarrow 2^{2^{G}}$ is the characteristic function of the game which for every coalition $C \subseteq A$ determines a set $V(C)$ of choices, the intended interpretation being that if $H \in V(C)$ then one of the choices is available to coalition C is to bring about all the goals in H simultaneously. $\rhd_i \subseteq G_i \times G_{i,}$ is a partial order over G_i representing *i's* preference relation, so that $g_r \rhd_i g_t$ indicates that i would rather have goal g_r satisfied than goal g_t

We say a set of goals H satisfies agent i if $H \cap G_i \neq \Phi$. We say that H satisfies $C \subseteq A$ if it satisfies every member of C. Also, we say that H is feasible for coalition C if $H \in V(C)$. The coalition preference which means the coalition likes one goal set more than the other goals set. We introduce the following definition.

Definition 2 A coalitions, called by C, can achieve all the goals sets of it and every members of it are satisfied in that sets.

$$X(C) = \{G' \subseteq G: (\wedge_{i \subseteq C}(G' \cap G_i \neq \Phi) \wedge (G' \in V(C))\},$$

$$\text{let } X^{\Gamma} = \cup_{C \subseteq A} X(C)$$

Definition 3 A coalition prefers a goals set than the other set of goals in X^{Γ} $C \subseteq A$, H, $H' \in X^{\Gamma}$, we say C strongly prefers the goal set *H to H', denoted as* $H \supset_c H'$ *if*

1 $H \in X(C)$,

2 $\forall i \in C$, $\exists r_i \in H \cap G_{i,}$ $\forall s_i \in H' \cap G_{i,}$ $r_i \rhd_i s_i$.

We say C weakly prefers the goal *H to H'*, denoted as $H \succ_c H'$ *if*

1 $H \in X(C)$,

2 $\forall i \in C$, $\forall s_i \in H' \cap G_{i,}$ $\exists r_i \in H \cap G_i$: $r_i \rhd_i s_i$

The $H \supset_c H'$ means that coalition C can achieve H which will satisfy every member of it and every member has a goal in $H \cap G_i$ which is better than any other goals from $H' \cap G_i$. $H \succ_c H'$ indicates that for $\forall i \in C$, no matter which goals of $H' \cap G_i$, i will find a better goal in $H \cap G_i$. In this definition, H' is not required to be in $X(C)$; we use $\rhd_c$ to express $\supset_c$ or $\succ_c$ in the following paper

The following example is concrete description of the coalition with preference

Example1 let Γ_1 be a Qualitative Coalitional Games with Preferences (QCGPs)

$\Gamma_1=<A,\ G\ ,G_1,\ G_2\ ,G_3,\ V,\ \rhd_1,\ \rhd_2, \rhd_3>$, *where*

$A=(a_1,\ a_2,\ a_3)$, $G=(g_1\ ,g_2\ ,g_3,\ g_4,\ g_5.\ g_6)$

$G_1=(g_1,\ g_4)$: $g_1 \rhd_1 g_4$

$G_2=(g_2,\ g_3,\ g_4)$: $g_2 \rhd_2 g_3 \rhd_2 g_4$

$G_3=(g_4,\ g_5,\ g_6)$: $g_4, \rhd_3 g_5 \rhd_3 g_6$

$V(a_1)=\{(g_1,g_2)\}$:

$V(a_2,a_3)=\{(g_1,g_3),\ g_6\}$:

$V(a_1,a_3)=\{(g_4,\ g_5),\ (g_1,\ g_4),\ (g_1,\ g_6)\}$:

We will make use of Γ_1 in later example

III. The Logic for QCGPs

Logic for expressing properties of individual of QCGs has been given in [9]; the key idea is that the language is defined in two parts: *Lc* is the satisfaction language, and is used to express properties of choices made by agents. The basic constructs in that language are of the form sat_i, meaning "agent i is satisfied". The overall language *L(QCGs)* is used for expressing properties of QCGs themselves. The main construct in that language is of the form $<C>\varphi$, where φ is a formula of the satisfaction language, and means that C have a choice such that this choice makes φ true. For example, $<3>(sat_1 \wedge sat_4)$ will mean that agents *3* has a choice that simultaneously satisfies agents *1* and *4*, we add preferences to that language. So we improve the logic by adding preference operations

A. The formal expression of logic

Formally, the grammar φ_c defines the satisfaction language *Lc*, while φ_q defines the QCGPs language *L(QCGPs)*.

$\varphi_c ::= sat_i \mid \neg\varphi_c \mid \varphi_c \vee \varphi_c$

$\varphi_q ::= <C>\varphi_c \mid [C]\varphi_c \mid \rhd_C \varphi_c \mid \neg\varphi_q \mid \varphi_q \vee \varphi_q$

Here, $i \in A$, $C \subseteq A$, the other propositional connectives ($\wedge$, $\rightarrow$, $\leftrightarrow$) is also used in the language of *Lc* and *L(QCGPs)*, the $[C]\varphi$ means that no matter what strategy C take, φ will be true and can be written as $\neg<C>\varphi$.

When $\Gamma=<A,\ G\ ,G_1,\ldots,G_n, V,\ \rhd_1,\ldots,\ \rhd_n>$ is a QCGPs, $H \subseteq G$, and $\varphi \in Lc$, $\Gamma.\ H \models_Q \varphi$ is defined as fellows.

$\Gamma.\ H \models_Q sat_i$ iff $G_i \cap H \neq \Phi$

$\Gamma.\ H \models_Q \neg sat_i$ iff not $\Gamma.\ H \models_Q sat_i$

$\Gamma.\ H \models_Q \varphi_1 \vee \varphi_2$ iff $\Gamma.\ H \models_Q \varphi_1$ or $\Gamma.\ H \models_Q \varphi_2$

When φ is *L(QCGPs)* formula, $\Gamma \models_Q \varphi$ is defined as felloows

$\Gamma \models_Q <C>\psi$ iff there is a $H \in V(C)$, such that $\Gamma.\ H \models_Q \psi$

$\Gamma \models_Q \rhd_C \psi$ iff there is a $H \in V(C)$, $H' \in X^{\Gamma}$, $\Gamma.\ H \models_Q \psi$, $\Gamma.\ H' \models_Q \psi$ and $H \rhd_C H'$

$\Gamma \models_Q \neg\psi$ iff not $\Gamma \models_Q \psi$

$\Gamma \models_Q \psi_1 \vee \psi_2$ iff $\Gamma \models_Q \psi_1$ or $\Gamma \models_Q \psi_2$

The preference of coalition of C means C has better choice to satisfy their goals. We will use the logic in section 4 to Branching Temporal framework.

Example 2: let Γ_1 be as in Example 1 Then

$\Gamma_1 \models_Q (a_1)(sat_1 \wedge sat_2\ )$

$\Gamma_1 \models_Q (a_2,a_3)sat_1 \wedge (a_2,a_3)sat_2) \wedge \neg((a_2,a_3)(sat_1 \wedge sat_2))$

$\Gamma_1 \models_Q \neg\ ((a_1,a_3)\ sat_2)$

$\Gamma_1 \models_Q \rhd\ (a_1,a_3)\ sat_1$

B. Expressive power of L(QCGPs)

The Expresive power of *L(QCGs)* is given in [9] by analyzing some properties of it, we pay attention to the properties of the preference which means what *L(QCGPs)* can express is that coalition can prefer some set of agents concurrently, we are not interested in neither how and why the coalitions prefer some goals, nor why an agent prefer one goal to the other one. We will use QCGPs-simulation to show the properties of preference, In other words, the language can not differentiate the preference of two games Γ and **Γ'** iff QCGPs-simulate each other

A relation

$$Z \subseteq \bigcup_{c \in A} V(C) \times V'(C)$$

is a QCGPs-simulation between two QCGPs, Γ=<A, G , $G_1,\ldots,G_n, V, \rhd_1,\ldots, \rhd_n$>, and $\boldsymbol{\Gamma'}$= <A, G' ,$G_1',\ldots, G_n'$, V' ,$\rhd_1',\ldots,\rhd_n'$> iff the following conditions hold for all coalitions C.

1 if $HZ\boldsymbol{H}$ then if $H \cap G_i = \Phi$, iff $\boldsymbol{H} \cap G_i' = \Phi$ for all i (the satisfaction condition), if $H \cap G_i \neq \Phi$ and there is $H' \in X^{\Gamma}$, $H \rhd_i H'$ iff $\boldsymbol{H} \cap G_i' \neq \Phi$ and there is $\boldsymbol{H'} \in X^{\Gamma'}$, $\boldsymbol{H} \rhd_i \boldsymbol{H'}$ for all i (the preference condition),

2 for every $H \in V(C)$, there is a $\boldsymbol{H} \in V'(C)$ such that $HZ\boldsymbol{H}$ (Z is total)

3 for every $\boldsymbol{H} \in V'(C)$, there is a $H \in V(C)$, such that $HZ\boldsymbol{H}$ (Z is surjective)

If there is a QCGP-simulation between Γ and $\boldsymbol{\Gamma'}$, we write $\Gamma \rightleftharpoons \boldsymbol{\Gamma'}$

Example 3: Let Γ_2=<A, G' ,G_1', G_2', G_3', V',$\rhd_1'$, $\rhd_2'$,$\rhd_3'$ > be the *QCG* with the same agents as in Γ_1

$A=(a_1, a_2, a_3)$, $G'=(f_1, f_2, f_3, f_4,)$

$G_1'=(f_1, f_3)$: $f_1 \rhd_1 f_3$

$G_2'=(f_1, f_4)$: $f_1 \rhd_2 f_4$

$G_3'=(f_2, f_3\ f_4,)$: $f_3 \rhd_3 f_4, \rhd_3 f_2$

$V(a_1)=\{(f_1)\}$:

$V(a_2,a_3)=\{(f_1), (f_2)\}$:

$V(a_1,a_3)=\{(f_1,f_3)\}$:

Then $\Gamma_1 \rightleftharpoons \Gamma_2$.The relation *Z consisting of the following pairs is a QCGP-simulation between* Γ_1 *and* Γ_2

$\{(g_1,g_2), (f_1)\}$: $\{(g_1,g_3), (f_1)\}$: $\{(g_1, g_4), (f_1,f_3)\}$: $\{g_6), (f_2)\}$

Note that Z not a function, nor the inverse of a function.

We can simulate any choice in one game with a choice in the other, and vice versa, if there is QCGPs-simulation for two games.

We write $\Gamma \equiv \boldsymbol{\Gamma'}$ iff $\forall_{\varphi \in L(QCGPs)}[\Gamma.\ H \models_Q \varphi \Leftrightarrow \Gamma'.\ \boldsymbol{H} \models_Q \varphi]$

Theorem 1.The preference is invariant under QCGPs-simulation: $\Gamma \rightleftharpoons \boldsymbol{\Gamma'} \Rightarrow \Gamma \equiv \boldsymbol{\Gamma'}$

Proof: let Γ=<A, G ,$G_1,\ldots,G_n$, V, $\rhd_1,\ldots,\rhd n$> and $\boldsymbol{\Gamma'}$=<A, G' ,$G_1',\ldots,G_n'$, V',$\rhd_1',\ldots,\rhd_n'$> with $\Gamma \rightleftharpoons \boldsymbol{\Gamma'}$, first, we show that

$$HZ\boldsymbol{H} \Rightarrow (\Gamma.\ H \models_Q \psi \Leftrightarrow \boldsymbol{\Gamma'}.\ \boldsymbol{H} \models_{Q'} \psi) \qquad (1)$$

For any ψ by induction over ; For the satisfaction condition, $\Gamma.\ H \models_Q \psi$ iff $H \cap G_i \neq \Phi$, iff, $\boldsymbol{H} \cap G_i \neq \Phi$ iff $\boldsymbol{\Gamma'}.\ \boldsymbol{H} \models_Q \psi$. For the preference condition, there is $H' \in X^{\Gamma}$, $\Gamma.\ H \models_Q \psi$, $\Gamma.\ H' \models_Q \psi$, $H \rhd_i H'$ iff $\boldsymbol{H'} \in X^{\Gamma'}$, $\boldsymbol{\Gamma'}.\ \boldsymbol{H} \models_Q \psi$, $\boldsymbol{\Gamma'}.\ \boldsymbol{H'} \models_Q \psi$, $\boldsymbol{H} \rhd_i \boldsymbol{H'}$. The inductive step (negation and disjunction) is straightforward. We now show that

$$\Gamma.\ \models_Q \varphi \Leftrightarrow \boldsymbol{\Gamma'}.\ \models_Q \varphi$$

for any φ by induction on φ For the base case, let $\varphi = \rhd_C \psi$, for the direction to the right, if $\Gamma.\ H \models_Q \psi$ then there is a $H \in V(C)$, $H' \in X^{\Gamma}$, such that $\Gamma.\ H \models_Q \psi$, Γ, $H' \models_Q \psi$ and $H \rhd_C H'$ by totality of Z, there is a $\boldsymbol{H} \in V'(C)$ such that $HZ\boldsymbol{H}$**,** by (1) then $\boldsymbol{\Gamma'}.\ \boldsymbol{H} \models_Q \psi$ and $\boldsymbol{H'} \in X^{\Gamma'}$, $\boldsymbol{\Gamma'}.\ \boldsymbol{H'} \models_Q \psi$, $\boldsymbol{H} \rhd_i \boldsymbol{H'}$. The direction to the left is symmetric: if $\boldsymbol{\Gamma'}.\ \boldsymbol{H} \models_Q \psi$, there is $\boldsymbol{H} \in V'(C)$, $\boldsymbol{H'} \in X^{\Gamma'}$, $\boldsymbol{\Gamma'}.\ \boldsymbol{H} \models_Q \psi$, $\boldsymbol{\Gamma'}.\ \boldsymbol{H'} \models_Q \psi$, $\boldsymbol{H} \rhd_i \boldsymbol{H'}$ by surjective of Z, there is a $H \in V(C)$ such that $HZ\boldsymbol{H}$**,** by (1) then $\Gamma.\ H' \models_Q \psi$ and $H' \in X^{\Gamma'}$, $\Gamma.\ H \models_Q \psi$, $H \rhd_i H'$, the step (negation and disjunction) is straightforward. The proof of invariant of satisfaction is given in [9] ■

Theorem 2. Let Γ, Γ' be defined over the same set of agents: $\Gamma \rightleftharpoons \boldsymbol{\Gamma'} \Leftarrow \Gamma \equiv \boldsymbol{\Gamma'}$

Proof Let Γ=<A, G, $G_1,\ldots,G_n$, V, $\rhd_1,\ldots,\rhd_n$> and $\boldsymbol{\Gamma'}$=<A, G' ,$G_1',\ldots,G_n'$, V',$\rhd_1',\ldots,\rhd_n'$> with $\Gamma \equiv \boldsymbol{\Gamma'}$, with any coalition C and any choice $H \in V(C)$, associate the set S_H^C={i: $H \cap G_i \neq \Phi$, $\exists H' \in X^{\Gamma}$, $H \rhd_i H'$} of agents satisfied if C prefers H to H'. Similarly for $\boldsymbol{\Gamma'}$: $T_{\boldsymbol{H}}^{C'}$ ={i: $\boldsymbol{H} \cap G_i' \neq \Phi$, $\exists \boldsymbol{H'} \in X^{\Gamma'}$, $\boldsymbol{H} \rhd_i \boldsymbol{H'}$}. For any $\boldsymbol{H} \in V'(C)$. We define a QCGPs- simulation Z: $\Gamma \rightleftharpoons \boldsymbol{\Gamma'}$ as follows: for every coalition C and pair of choices $H \in V(C)$, $\boldsymbol{H} \in V'(C)$,

$$HZ\boldsymbol{H} \Leftrightarrow S_{H'}^{C} = T_{\boldsymbol{H}}^{C'}.$$

We must show that Z is total, i.e., that if $H \in V(C)$ then there is $\boldsymbol{H} \in V'(C)$, such that $S_{H'}^{C} = T_{\boldsymbol{H'}}^{C'}$. Suppose not: assume that $i \in S_{H'}^{C}$ and $i \notin T_{\boldsymbol{H}}^{C'}$, for all $\boldsymbol{H} \in V'(C)$ Then $\Gamma.\ \models_Q \rhd_C sat_i$, and $\boldsymbol{\Gamma'}.\ \models_Q \neg \rhd_C sat_i$ which contradicts the fact that $\Gamma \equiv \boldsymbol{\Gamma'}$ the same is to $i \notin S_{H'}^{C}$, $i \in T_{\boldsymbol{H'}}^{C'}$. Similarly, we must show that Z is surjective, the proof is the same as Z is total. Finally, we show that the satisfaction condition holds if. $HZ\boldsymbol{H}$**,** then $H \cap G_i \neq \Phi$ iff $i \in S_{H'}^{C}$, iff, by the definition of Z, $i \in T_{\boldsymbol{H'}}^{C'}$ iff $\boldsymbol{H} \cap G_i' \neq \Phi$ ■

C. Axiomatisation for QCGPs

We give the axiomatisation of qualitative coalitional games, and show its soundness and completeness. We use K(QCGPs) to express the axiomatisation for QCGPs for close resemblance to the modal system K, which also indicates that our logic, is in a sense, a weakest basic system for QCGPs, The system K(QCGPs) over the language *L(QCGPs)* is defined as follows, where φ,ψ are arbitrary *L(QCGPs)* formulae, α,β are arbitrary *Lc* formulae and *C* is an arbitrary coalition:

Prop If φ is an *L(QCGPs)*-instance of a propositional tautology, then φ is provable

K $[C](\alpha \rightarrow \beta) \rightarrow ([C]\alpha \rightarrow [C]\beta)$ is provable. $\rhd_C(\alpha \rightarrow \beta) \rightarrow (\rhd_C \alpha \rightarrow \rhd_C \beta)$ is provable

MP If $\varphi, \varphi \rightarrow \psi$ are provable, then ψ is provable

Nec If α is an (*Lc*) instance of a propositional tautology, then $[C]\alpha$, $\rhd_C \alpha$ are provable

It is easy to see that the deduction theorem holds for K(QCGPs).We will need the following properties of

K(QCGPs). The proofs are straightforward for readers familiar with modal logic

LEMMA 1. $\alpha,\beta\in Lc$:

1 $\models_{\mathrm{K(QCGP)}} \langle C\rangle(\alpha\wedge\beta) \rightarrow \langle C\rangle(\alpha)$

2 $\models_{\mathrm{K(QCGP)}} \langle C\rangle(\alpha\vee\beta) \rightarrow (\langle C\rangle(\alpha) \vee \langle C\rangle(\beta))$

3 $\models_{\mathrm{K(QCGP)}} (\langle C\rangle(\alpha)\wedge[C](\alpha\rightarrow\beta)) \rightarrow \langle C\rangle(\beta)$

4 $\models_{\mathrm{K(QCGP)}} \rhd_C(\alpha\wedge\beta) \rightarrow \rhd_C\alpha$

5 $\models_{\mathrm{K(QCGP)}} \rhd_C(\alpha\vee\beta) \rightarrow (\rhd_C\alpha\vee\rhd_C\beta)$

Theorem 3 (SOUNDNESS & COMPLETENESS)

For any $\Omega\subseteq L(QCGPs)$, $\varphi\in L(QCGPs)$,

$$\Omega\models_Q\varphi \Leftrightarrow \Omega\models_{\mathrm{K(QCGPs)}}.\varphi$$

Proof: For soundness (the direction to the left), it is easy to see that the axioms are valid, and that the rules preserve logical consequence. For completeness, let $\psi\subseteq L(QCGPs)$ be K(QCGPs) consistent. We show that ψ is satisfied by some QCGPs. Let A be the set of agents and let n = $|A|$,. Let Δ be a $L(QCGPs)$ maximal and K(QCGPs) consistent set containing ψ. We now construct $\Gamma=\langle A, G, G_1,\ldots,G_n, V, \rhd_1,\ldots,\rhd_n\rangle$ intended to satisfy ψ. as follows:

G = { sat$_1$, . . . , sat$_n$}

G_i = { sat$_{n,}$ sat$_i$ } sat$_n\rhd_i$sat$_i$ for each i

$H\in V(C) \Leftrightarrow \rhd_c\mathcal{E}H\in\Delta$ for any $H\in G$ where

$\mathcal{E}H= \bigwedge_{sat_i\in H, sat_n\in H} \mathrm{sat}_i \wedge \bigwedge_{i\in A, sat_i\notin H} \neg\mathrm{sat}_i$, we show that

$$\text{Gamma} \models_Q \gamma \Leftrightarrow \gamma\in\Delta$$

For any γ by structural induction over γ For the base case, $\gamma=\rhd_c\alpha$, $\alpha\in L_{c,}$ Again, we use induction on the structure of α, For the (nested) base case, let α=sat$_i$, For the direction to the right, if $\Gamma \models_Q \gamma$ then there is an $H\in V(C)$, and $\exists\boldsymbol{H}\in X^{\Gamma}$ and $\Gamma.\ H \models_Q \gamma$, $\Gamma.\ \boldsymbol{H} \models_Q \gamma$, $H\rhd_c\boldsymbol{H}$, then sat$_i\in H$, and by Lemma 1.4, $\gamma=\rhd_c\mathrm{sat}_i\in\Delta$. For the direction to the left, Let $\rhd_c\mathrm{sat}_i\in\Delta$, $S\subseteq G$, $S'\subseteq G$ let

$$X_i= \bigvee_{S\subseteq G} \mathcal{E}H(S\cup sat_i, sat_n)$$

$$\mathbf{X_i}'= \bigvee_{S'\subseteq G} \mathcal{E}\boldsymbol{H}(S'\cup sat_i).$$

sat$_i\rightarrow$X$_i$, sat$_i\rightarrow\mathbf{X_i}'$ is a Lc instance of a propositional tautology, by Nec $\rhd_c(\mathrm{sat}_i\rightarrow X_i)\in\Delta$, $\rhd_c(\mathrm{sat}_i\rightarrow\mathbf{X_i}')\in\Delta$, then $\exists S,S'\ \Gamma.(S\cup\{\mathrm{sat}_i, \mathrm{sat}_n\})\models_Q\gamma$, $\Gamma.(S'\cup\{\mathrm{sat}_i\})\models_Q\gamma$, and $(S\cup\{\mathrm{sat}_i, \mathrm{sat}_n\})\rhd_c(S'\cup\{\mathrm{sat}_i\})$, so $\Gamma \models_Q\rhd_c\mathrm{sat}_i$ which concludes the proof of the direction to the left in the innermost induction proof. Both the inner and the outer induction steps (negation and disjunction) are straightforward. The proof about $\gamma=\langle C\rangle\alpha$ is be given in [9]■

Ⅳ. BRANCHING TEMPORAL QCGPS

Using temporal method is an effective ways for repeated games, to CTL, at each node of it, a (possibly different) QCGPs Γ is played. A Branching temporal qualitative coalitional games (CTQCGPs) is then a triple M =<S, R, Q> where:

S: is a set of states;

R: is a total binary relation $R\subseteq S\times S$. i.e.. $\forall s\in S$, $\exists t\in S$, $\langle s, t\rangle\in R$, and

Q: S→Q, where Q is the class of all QCGPs, is a function associating a qualitative coalitional games Q(S) $=\langle A, G, G_1,\ldots,G_n, V, \rhd_1,\ldots\rhd_n\rangle$with every state $s\in S$.

A. A Logic for CTQCGPs

In this section we provide the formal syntax and semantics for representative systems of Branching Temporal QCGPs propositional by branching time temporal logics, CTL (Computational Tree Logic) allows basic temporal operators of the form: a path quantier–either A (for all futures pathe) or E (for some future path) followed by a single one of the usual linear temporal operators G (always), F (sometime), X (next time) or U (until).Formally, the language of *L(CTQCGPs)* is defined by the grammar φ_t

$$\varphi_t ::= \langle C\rangle\varphi_t|\neg\varphi_t|\rhd_c\varphi_t\ |\varphi_t\vee\varphi_t\ |\ E(\varphi_t U\psi_t)\ |\ EX\varphi|\ EG\varphi|$$

The remaining temporal operators to express eventuality and universality can be derived in standard way, for instance: $EF\varphi_t = E(\mathbf{T}\cup\varphi_t)$, and $AG\varphi_t = \neg EF\neg\varphi_t$. CTL formulae are interpreted in *Kripke* models. When M =<S, R, Q>is a CTQCGPs，A path $\pi =\langle\pi_0, \pi_1, \pi_2, \cdots\rangle$ of M is an infinite sequence of states in s such that $(\pi_i, \pi_{i+1})\in R$ for all $i\geq 0$. $s\in S$, φ is a $L(QCGPs)$ formula, the satisfaction relation. M.s $\models_T \varphi$, is defined as follows (the cases for negation and disjunction are defined as usual).

M. s $\models_T \varphi$ iff Q(s) $\models_Q \varphi$ when $\varphi\in L(QCGPs)$

M. s $\models_T E(\varphi_t U\psi_t)\varphi$ iff there exists a path π such that π_0 =s and a $k\geq 0$. Such that $\pi_k \models_Q \psi$ and $\pi_i \models_Q \varphi$ for all $0\leq i< k$, $Q(\pi_i)\models_Q \varphi$, φ, $\psi\in L(QCGPs)$.

M. s $\models_T EX\varphi$ iff there exists a path π such that π_0 = s and $Q(\pi_1) \models_Q \varphi$, $\varphi\in L(QCGPs)$

M. s $\models_T EG\varphi$ iff there exists a path π such that π_0 = s and $Q(\pi_i) \models_Q \varphi$ for all $i\geq 0$.

We will henceforth use *L(CTQCGPs)* to refer to both the language, and the logic we have defined over this language.

B. Expressive Power of CTQCGPs

The notion of simulation for QCGPs (Section 2.2) can be naturally used to the branching temporal case. When M = (S, R, Q) and M'= (S', R', Q') are CTQCGPs and $s\in S$, $s'\in S'$ we define.

$M.s \rightleftharpoons_T M'.s' \Leftrightarrow Q(s) \rightleftharpoons Q'(s')$,

$M. \rightleftharpoons_T M'$, $\Leftrightarrow \forall s, \exists s'\ M.s\rightleftharpoons_T M'.s'$, and $\forall s', \exists s$, $M.s\rightleftharpoons_T M'.s'$

The notion of elementary equivalence for CTQCGPs over the language *L(CTQCGPs)* can be defined as follows. $M.s\equiv M'.s'$ iff, for every $\varphi\in L(CTQCGPs)$, M. s $\models_T \varphi$ iff M' s' $\models_T \varphi$ $M\equiv M'$, iff for $\forall S'$, $\exists S$, $M.s\equiv_T M'.s'$ and $\forall S'$, $\exists S$, $M.s\equiv M'.s'$. Note that in the branching temporal case, the fact that M,s $\rightleftharpoons_T$ M',s' is not sufficient for M,s≡M',s' to hold.

C. Satisfiability

The satisfiability problem for L(CTQCGPs) is as follows: given a formula $\varphi \in$ *L(CTQCGPs)*, does there exist a *CTQCGPs* M and s∈S, such that M. s $\models_T \varphi$?

Theorem 4. The sat. probl. for L(CTQCGPs) is a PSPACE-complete problem.

The satisfiability problem of TQCGs in [9] is given by LTL+K$_m$(the fusion of LTL and multi-modal K),so the same method is for CTL+K$_{m,}$ in [12,13]pointing that the complexity of CTL+K$_m$ is PSPACE-complete problem. The detail context to translate *L(CTQCG)* formula to LTL+K$_m$ is in [9], so the same method can be used for CTL+K$_m$

D. Satisfiability

The system K(CTQCGPs) over the language *L(CTQCGPs)* is defined as follows, where φ,ψ are arbitrary *L(CTQCGPs)* formulae, *A,B* are arbitrary *L(QCGPs)* formulae, α β are arbitrary *Lc* formulae and *C* an arbitrary coalition. For simplicity, we write T instead of K(TQCGPs) for derivability in K(CTQCGPs).

Prop If *A* is an *L(QCG)* instance of a propositional tautology, then $\vdash_T A$

K $\vdash_T [C](\alpha \to \beta) \to [C](\alpha) \to [C](\beta)$

MP if $\vdash_{K(QCGPs)} A$ and $\vdash_{K(QCGPs)} A \to B$, then $\vdash_T B$

Nec If α is an (*Lc*) instance of a propositional tautology, then $\vdash_T [C]\alpha$, $\vdash_T \rhd_c \alpha$

A1 $EX(\varphi \vee \psi) \leftrightarrow (EX\varphi \vee EX\psi)$

A2 $AX\varphi \leftrightarrow \neg\, EX\neg\varphi$

A3 $AG(\varphi \to (\neg\psi \wedge EX\varphi)) \to (\varphi \to AF\psi)$

A4 $AG(\varphi \to (\neg\psi \wedge AX\varphi)) \to (\varphi \to EF\psi)$

A5 $AG(\varphi \to \psi) \to (EX\varphi \to EX\psi)$

U1 $E(\varphi U\psi) \leftrightarrow (\psi \vee (\varphi \wedge EX(\varphi U\psi)))$

U2 $A\ (\varphi U\psi) \leftrightarrow (\psi \vee (\varphi \wedge AX\ (\varphi U\psi)))$

Prop if φ is an (*L(CTQCGPs)*) instance of a propositional tautology, then $\vdash_T \varphi$

Nec If $\vdash \varphi$ then $\vdash AG\varphi$

MP If $\vdash \varphi$ and If $\vdash \varphi \to \psi$ then $\vdash \psi$

Axioms Prop and K and rules MP and Nec say that every K(CTQCGPs)-theorem is also a K(CTQCGPs)-theorem. The subsystem consisting of axioms A1–U2 and rules Prop–Nec is a version (with *L(QCGPs)* formulae in place of atomic propositions) of an axiomatisation of branching time logic proved to be sound and complete in [10].

Theorem 5 (SOUNDNESS & COMPLETENESS). For any $\varphi \in$ *L(TQCGPs)* $\vdash_T \varphi \Leftrightarrow \models_T \varphi$

Proof: The logic K(CTQCGPs) is a temporalisation of K(QCGPs): the language of K(CTQCGPs) has atomic K(QCGPs) formulae in place of atomic propositions; the semantic structures of K(CTQCGPs) identifies a semantic structure for K(QCGPs) by using *Kripke* model for interpreting K(QCGPs) formulae; and the rules of K(CTQCGPs) are the rules of the temporal logic for temporal formulae in addition to axioms/rules of K(QCGPs) formulae. Finger [11] show that the temporalisation of a sound and complete system is sound and complete. The theorem thus follows immediately from Theorem 3. ■

E. An Example

We add Branching Temporal dimension with preference to the example given in [9]. There are there agents and server providing web service, all agents needs to access the server from time to time, There are three basic action for agents, read access, write access and wait, For the integrity, web service is violated if two agents write access are granted (inconsistent writes) or read and write access are granted (inconsistent reads), or no action for any agents (inefficiency) at the same time

Let M =<S, R, Q> be a CTQCGPs where S is some infinite set of states, and R is total binary relation over S×S and Q are holds for Q(s: s∈S)=<$A, G, G_1, G_2, G_3, G_{ser}, V, \rhd_1, \rhd_2\ \rhd_3\ \rhd_{ser}$>

A={1,2,3,ser}, We model the agents as players 1, 2 and 3, and the server as player ser

G={$s_1,s_2,s_3,s_4,s_5,s_6,s_7$*), That each of these goals are achieved means that right now.*

s_1*: every agent is granted read access*

s_2*: agent 1 is granted write access*

s_3*: agent 2 is granted write access*

s_4*: agent 3 is granted write access*

s_5*: Server prefer write access agent 1 to agent 2*

s_6*: Server prefer write access agent 1 to agent 3*

s_7*: Server prefer write access agent 2 to agent 3*

for the preference of every agent, for example, players 1 more like *(s1, s2)* than *(s1)*, Notice that when players 1 get the goal (s1, s2) that means he can read or write, not that read and write at the same time.

$G_1 = \{(s_1), (s_1, s_2)\} \quad (s_1, s_2) \rhd_1 (s_1)$

$G_2 = \{(s_1), (s_1, s_3)\} \quad (s_1, s_3) \rhd_2 (s_1)$

$G_3 = \{(s_1), (s_1, s_4)\} \quad (s_1, s_4) \rhd_3 (s_1)$

$G_{ser} = \{(s_1), (s_5), (s_6), (s_7)\}$

It is assumed that for *ser,* its preference will be changed with time: Given the symbol t (minute) for time with the state changes, *ser* has different preference.

If t mod 2 = 0 $(s_5) \rhd_{ser} (s_6)\ \rhd_{ser} (s_1)$

If t mod 3 = 0 $(s_6) \rhd_{ser} (s_7)\ \rhd_{ser} (s_1)$

If t mod 5 = 0 $(s_7) \rhd_{ser} (s_5)\ \rhd_{ser} (s_1)$

otherwise $(s_1)\ \rhd_{ser}\ (s_5) \rhd_{ser} (s_7) \rhd_{ser}\ (s_6)$

$V(1, ser)=\{(s_1), (s_1, s_2)\},$

$V(2, ser)= \{(s_1), (s_1, s_3)\}$

$V(3, ser)= \{(s_1), (s_1, s_4)\}$

$V(1,2,ser)=\{(s_1), (s_1, s_2, s_5)\},$

$V(2,3,ser)=\{(s_1), (s_1, s_3, s_6)\},$

$V(1,3,ser) =\{(s_1), (s_1, s_4, s_7)\},$

$V(1,2,3,ser) =\{(s_1,s_2,s_3,s_4,s_5,s_6,s_7)\},$

The following properties hold in the system

1 EF<*ser*>($sat_1 \wedge sat_2\ \wedge sat_3$) all agents will be satisfied in sometime at some future path just like t=30

2 AG <*ser*>($sat_1 \vee sat_2 \vee sat_3$) for all future path, there is at least a agent will be satisfied

3 E [<*ser*>($sat_1 \vee sat_3$) U <*ser*>(sat_2)] agent 1 or 3 will be satisfied in some future path until agent 2 is satisfied.

Ⅴ. CHARACTERIZING CTQCGPS S

In this section, we investigate the axiomatic characterizations of various classes of CTQCGPs., We pay attention to the preference characteristic with time. As usual, when saying that a formula scheme' characterizes a property *P* of models, we mean that φ is valid in a model M iff M has property *P*; if only the right-to-left part of this biconditional holds, then we say property *P* implies φ.

A. Basic Correspondences

Let $h^s(C)$ denote the set of all agents that could possibly be satisfied (not necessarily jointly) by coalition *C* in state s:

$$h^s(C) = \{ i: i \in A \;\&\; \exists H \in V^s(C), G^s_i \cap H \neq \Phi \}$$

The "h" here is for "happpiness": We regard $h^s(C)$ as all the agents that *C* could possibly make happy in s. Thus the semantic property $i \in h^s(C)$ is a counterpart to the syntactic expression $<C>sat_i$ [9], we use $hp^s(C)$ to denotes the $h^s(C)$ with preference.

$hp^s(C) = \{i: i \in A \;\&\; \exists H \in X^s(C), \forall \boldsymbol{H} \in X^s(C)/G$, there is no such that case $(\boldsymbol{H} \cap G_i) \rhd_i (H \cap G_i)\}$ which means *C* have realized a preference set of goals in s.

$\forall s \in S$, for all path from s,

$s \rightarrow s', (i \in hp^s(C)) \rightarrow (i \in hp^{s'}(C))$ (AXHP)

if means *C* have realized a preference set of goals.in all the state immediately following s.

$\forall s \in S$, for a path from s,

$s \rightarrow s', (i \in hp^s(C)) \rightarrow (i \in hp^{s'}(C))$ (EXHP)

if means *C* have realized a preference set of goals in a state immediately following s.

Lemma 2. $\rhd_C sat_i \rightarrow AX \rhd_C sat_i$ characterizes AXHP, $\rhd_C sat_i \rightarrow EX \rhd_C sat_i$ characterizes EXHP

We also characterise the unperference property

$\forall s \in S$, for all state immediately following s,

$s \rightarrow s', (i \notin hp^s(C)) \rightarrow (i \notin hp^{s'}(C))$ (AXUP)

$\forall s \in S$, for a state immediately following s,

$s \rightarrow s', (i \notin hp^s(C)) \rightarrow (i \notin hp^{s'}(C))$ (EXUP)

Lemma 3. $\neg \rhd_C sat_i \rightarrow AX \neg \rhd_C sat_i$ characterizes AXUP , $\neg \rhd_C sat_i \rightarrow EX \neg \rhd_C sat_i$ characterizes EXUP,

To the time of future, eventually, *C* will be able to make *i* happy with preferences.

for all path $\exists s \in S, (i \in hp^s(C))$ (AFEH)

for a path $\exists s \in S$ in the path $(i \in ph^s(C))$ (EFEH)

for all path $\exists s \in S, (i \notin hp^s(C))$ (AFEU)

for a path $\exists s \in S$ in the path $(i \notin hp^s(C))$ (EFEU)

Lemma 4 $AF \rhd_C sat_i$ characterizes AFEH, $EF \rhd_C sati$ characterizes EFEH

$AF \neg \rhd_C sat_i$ characterizes AFEU $AF \neg \rhd_C sat_i$ characterizes EFEU

Finally, we consider safety properties. Which means *C* always have realized a preference goals set and *C* never can have realized a preference goals set

for all path $\forall s \in S, (i \in hp^s(C))$ (AGPH)

for a path $\forall s \in S$, in the path $(i \in hp^s(C))$ (EGPH)

for all path $\forall s \in S, (i \notin hp^s(C))$ (AGPU)

for a path $\forall s \in S$ in the path , $(i \notin hp^s(C))$ (EGPU)

Lemma 5 $AG \rhd_C sat_i$ characterizes AGPH, $EG \rhd_C sat_i$ characterizes EGPH

$AG \neg \rhd_C sat_i$ characterizes AGPU, $EG \neg \rhd_C sat_i$ characterizes EGPU

B. Basic Properties of preference Choice Sets

We consider whether a coalition has a preference set of goals in a state and whether it has a best choice.

Definition 4 for $C \subseteq A$, the maximal strongly preferred goal sets with respect to *C*, denoted $\mu^{\sqsupset}$ are defined through

$\mu^{\sqsupset}(C) = \{H \in X(C), \forall \boldsymbol{H} \in X(C)$, it is not the case that $H \sqsupset_c \boldsymbol{H}\}$ the maximal weakly preferred goal sets with respect to C, denoted

$\mu^{\succ}(C) = \{H \in X(C): \forall \boldsymbol{H} \in X(C)$, it is not the case that $\boldsymbol{H} \succ_c H\}$ In the event of $\mu^{\sqsupset}(C) = \mu^{\succ}(C)$ we write simply $\mu(C)$. To avoid excessive repetition, we use the relational symbol $\rhd$ to indicate either $\sqsupset$ or $\succ$

$\forall s \in S\ \mu_s^{\rhd}(C) = \Phi$， *C* never has preference choice.

$\forall s \in S\ \exists H \in \mu_s^{\rhd}(C) \neq \Phi$， $H \neq \Phi$, *C* has preference choice

$\forall s \in S\ \exists H \in X^s(C), \forall \boldsymbol{H} \in X^s(C)/H$， $H \rhd_C \boldsymbol{H}$**,** *C* has a best preference choice

C. Static preference Goal Sets and Choices

The goal sets with preference for each agent and the choice sets for each coalition are guaranteed to remain unchanged.for all path(existential quantifier is easy for reader to built)

$\forall s,s' \in S\ (H_i^{\,s} = H_i^{\,s}\ \&\ \rhd_i^{s} = \rhd_i^{s'})$ (ASGS)

the goal set with preference is static for agent *i*.

$\forall s,s' \in S\ (V(C)^s = V(C)^{s'}\ \&\ \rhd_i^{s} = \rhd_i^{s'})$ (ASC)

coalition *C*'s preference choices remain static

$\forall s,s' \in S$, $\mu(C)^s = \mu(C)^{s'}$) coalition C's maximal strongly preferred goal sets remain static

Lemma 6. Any model satisfying both ASGS and ASC also satisfies AXHP and AXUP, and as a consequence, ASGS and ASC together imply $\rhd_C sat_i \leftrightarrow AXPH \rhd_C sat_i$,

D. Dynamic preference Goal Sets of individual agent

Considering agent *i*'s goals set with preference is guaranteed to monotonically decrease over time. Roughly, this condition means that every agent is guaranteed to get no easier become to satisfy his preference over time. Formally a agent *i*'s preference goal sets in state s is better than all the immediately following s (existential quantifier is easy for reader to built)

$\forall s \in S$, for all path from s, $s \rightarrow s', \exists H \in \mu_s^{\rhd}(C)$,

$\forall \boldsymbol{H} \in \mu_{s'}^{\rhd}(C), \forall g_2 \in (\boldsymbol{H} \cap G_i), \exists g_1 \in (H \cap G_i), g_1 \rhd g_2$

The monotonically increasing over time is:

$\forall s \in S$, for all path from s, $s \rightarrow s'$, $\forall H \in \mu_s^{\rhd}(C)$,

$$\exists \boldsymbol{H} \in \mu_{s'}{}^{\rhd}(C),\ \exists g_2 \in (\boldsymbol{H} \cap G_i),\ \forall g_1 \in (H \cap G_i),\ g_2 \rhd g_1$$

E. Dynamic preference Choices

We also investigate the coalition's choice in time, which say that the sets of choices available to coalition C monotonically increase or decrease respectively. (existential quantifier is easy for reader to built)

$\forall s \in S$, for all path from s, s→s', $\mu_s{}^{\rhd}(C) \subseteq \mu_{s'}{}^{\rhd}(C)$ the coalition C' preferred goal sets are monotonically increase

$\forall s \in S$, for all path from s, s→s', $\mu_s{}^{\rhd}(C) \sqsupseteq \mu_{s'}{}^{\rhd}(C)$ the coalition C' preferred goal sets are monotonically decrease

F. Solution Concept

Definition 5 Let $\Gamma = <A,\ G\ ,G_1, \ldots, G_n,\ V,\ \rhd_1, \ldots, \rhd_n>$ For a coalition $C \subseteq A$ the core of C denoted $K^{\rhd}(C)$, is the set $\{\ H \in \mu^{\rhd}(C):,\ \forall C' \subset C,\ \forall \boldsymbol{H} \in \mu^{\rhd}(C'),$ it is not the case that $\boldsymbol{H} \rhd_C H\ \}$,then the coalition C is core sets.
To an agent, perhaps the key problem is whether at every time point there is some stable coalition, containing this agent.

stable(i)= $AG \bigvee_{C \in A, i \in C} K^{\rhd}(C)$, if there is a coalition satisfying the stable of agent i. which means, to agent i, there is a stable solution for it. More characteristics of QCGPs can be seen in [5]

VI. Conclusion

QCGs were introduced in [4] as a model for coalition games with non-numeric values payoff, it score how a coalition can be formed, to a agent. Although it has many goals as its desire, it can't be get all the goals in games, so preference goals is effective ways for an agent's strategy.

Formal coalition cooperating games is a hot point in social software and multi-agents systems, in this paper we introduce the problem from QCGPs and CTL, we investigate the basic concept and model of QCG with preference and give a logic language for it by using simplest modal logic K, and we talk about the expression power and axiomatisation of the logic. The CTL is used for the repeated games and some characteristic of CTQCGPs, such as realizing preference goals sets, best strategy of coalition, the stabilization of coalition and so on, are given

The further researches are multiple. First are the properties of CTQGCPs some principium investigations were made in [5], but considering the repeated games, more work should be taken. Second, temporalising QCGPs also has many ways, just like ATL and CTL*, The LTL and CTL are only reflects a simple case. The more complex temporal structure should be used for deeper percipience to coalition games. In addition, the dynamic goals and preferences are also important problems, in practice application, the goals and preferences of an agent are not static, for example, with the resource loss, a agent will decrease its goals and preferences or after by getting some goals, an agent will increase its goals for being unsatisfied in existing goals. So those factors will influence the structure and strategy of the coalition in games.

Acknowledgment

This work was supported by the 863 program under grant of 2007AA01Z425, 973 program under grant of 2007CB316502, the National Natural Science Foundation of China under grant of 60173026

References

[1] Marc Pauly, *Logic for Social Software*, Ph.D. Thesis, University of Amsterdam. ILLC Dissertation Series 2001-10, ISBN 90-6196-510-1.

[2] T. Sandholm, K. Larson, M. Andersson, O. Shehory, and F. Tohm´e. Coalition structure generation with worst case guarantees. *Artificial Intelligence*, vol 111(1–2):pp.209–238, 1999

[3] O. Shehory and S. Kraus. Methods for task allocation via agent coalition formation. *Artificial Intelligence*, vol 101(1-2): pp.165–200, 1998

[4] M. Wooldridge and P. E. Dunne. On the computational complexity of qualitative coalitional games. *Artificial Intelligence*, vol 158(1):pp.27–73, 2004

[5] P. E. Dunne and M. Wooldridge Qualitative coalitional reasoning with preferences. *Technical Report*, ULCS-04-008, Dept. of Computer Science, Univ. of Liverpool, March 2004.

[6] M.J. Wooldridge and P.E. Dunne. Preferences in qualitative coalitional games. Proc. Sixth Workshop on Game *Theoretic and Decision Theoretic Agents,* (GTDT'04), New York, ^{20}th July 2004, pp 29-38

[7] T. Sandholm. Distributed rational decision making. In G. Wei, editor, *Multiagent Systems*, pages 201–258. The MIT Press: Cambridge, MA, 1999

[8] M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. The MIT Press: Cambridge, MA, 1994.

[9] Thomas Ågotnes, Wiebe van der Hoek, Michael Wooldridge: Temporal qualitative coalitional games. AAMAS 2006: 177-184

[10] D. Gabbay, A. Pnueli, S. Shelah, and J. Stavi. On the temporal analysis of fairness. In POPL '80, pp. 163–173,1980. ACM Press

[11] M. Finger and D. M. Gabbay. Adding a temporal dimension to a logic system. *Journal of Logic, Language, and Information*, vol 1.pp.:203–233, 1992.

[12] Lomuscio, A. and Raimondi, F. The complexity of model checking concurrent programs against CTLK specifications. In Proceedings of the fifth international joint conference on Autonomous agents and multiagent system (AAMAS'06), pages 548 - 550, Hakodake, Japan. ACM Press

[13] Raimondi, F. and Lomuscio, A. The complexity of symbolic model checking temporal-epistemic logics. *In Proceedings of Concurrency, Specification & Programming (CS&P)*, pp 421–432. Warsaw University

Cheng Bailiang (1978-), male, Ph.D. candidate of computer science of tongji university, his research interest include Game logic, Trust computing, Parallel Computing

Zeng Guosu, (1964) male, 1964 received his BS, MS, and Ph.D. in computer software and application all from the Department of Computer Science and Engineering, Shanghai Jiaotong University. Nowadays, he is working in Tongji University as a professor, as well as a supervisor of Ph.D. candidates in computer software and theory. His research interests include parallel processing, heterogeneous computing, and grid theory. In the past years, he has taken charge of lots of research projects supported by national and local government. He has more than 60 papers published in national or international key journals. Some of them were cited by SCI or EI. He got award of Panwenyuan, an international education fund, and in 2004 he got the first prize of Liguohao for excellent teacher.

Jie Anquan(1975-**)**, male, instructor, master, College of Computer Information and Engineering, Jiangxi Normal University, his research interest include parallel， concurrency and Information Retrieval

A New Method to Improve the Maneuver Capability of AUV Integrated Navigation Systems

Zhen Guo
College of Automation, Harbin Engineering University, Harbin, China
Email: guozhenhrb@yahoo.com.cn

Yanling Hao and Feng Sun
College of Automation, Harbin Engineering University, Harbin, China

***Abstract*—The maneuver characteristic of the most commonly used AUV integrated navigation systems was investigated in this paper. After analyzing the error cause of conversional used Kalman filter of SINS/DVL integrated navigation systems in maneuver state, a novel method was proposed which is to use the output of complex navigation systems to revise the SINS in real-time, and an improved adaptive Kalman filter was discussed here to reach the seamless changing of the whole system. The measurement remnant method was introduced to judge whether the bearing change event happened or not. The whole design was aiming to reach the smooth transition between the different motion states and improve the maneuver capability of the AUV navigation system. The simulation results confirms the new approach could restrain the oscillation of Kalman filter in motion changing state and improve the accuracy of the AUV integrated navigation systems.**

***Index Terms*—maneuver, integrated navigation systems, adaptive Kalman filter, AUV**

I. INTRODUCTION

AUV has drawn a lot of attentions widely across the world because of its flexibility, small-dimension and multi-tasks characteristic which requires the navigation systems should have the long-term independently working capability and also provide accurate positioning information in the complicated marine environment. The precision of the navigation system can directly affect the quality of whole task because the most important factor for an AUV to successfully complete a typical survey mission is to follow a path specified by the operator as accurately as possible. Unlike the UAV (unmanned aerial vehicle), AUV presents a more challenging navigational problem, because it is always lack of the direct positioning signal from GPS (global positioning system) to revise the SINS positioning information when it is in submerging state and the normally used sensor DVL could only provide the velocity information. H. Stutters and H. Liu discussed different methods in different underwater environments of AUV and also addressed their current problems [1].

As is known to us all, the maneuver scenario is always an intractable and unavoidable problem not only for target tracking but also for navigation issue. As the conversional Kalman filter has limitation in such situation, a lot of researches have been made aiming to decrease the oscillation, speed up the convergence process with the help of GPS or an existing beacon to revise the position. The underwater maneuvering problem is still one of the most challenging scenarios of AUV, considering above requirements are hardly met during its mission.

One solution is to equip the AUV with more accurate acceleration and gyro sensors to somehow decrease the drift of SINS in long-term working range, but this always means higher cost and also limited by the mechanical technique.

The other solution is to use some methods to decrease the drift bias through rotation of SINS. S. Ishibashi, etc. proposed their method to improve the performance of INS which was put on a turntable with one rotational axis and rotated by it according to some rules [2].

The third solution is to redesign the integrated navigation system and improve the adaptive filtering algorithm to be adapted to the dynamic motion which is discussed later in this paper. B. Liu and J. Fang introduced a method to use the normalized observability of states as a factor to adaptively feedback the SINS with the estimated output value of Kalman filter [3]. But something should be done to improve the theory because the vehicle motion could improve the observability, it will also involve the short term oscillation of Kalman filter when switching between different state, such oscillation will even lead to the divergence of SINS if the estimated value is directly used to judge the SINS. Z. Guo and F. Sun proposed initial idea to adopt the dead reckoning algorithm using the heading information of the SINS and DVL to substitute the SINS/DVL integrated system [4] but without further investigation about the whole design of the changing method. The SINS/DVL integrated navigation system and the Compass/DVL dead reckoning system are the most common choice for AUV underwater working mode. Both systems have their own advantage in linear motion and curvilinear motion. SINS/DVL integrated navigation system normally uses

doi:10.4304/jcp.5.5.757-764

Kalman filter which has a bad performance in maneuvering state; considering that the SINS' low frequency noise and the sensor bias will lead to poor accuracy in long term, the close loop feedback will even cause the divergence of the filtering process at the motion changing time. The Compass/DVL dead reckoning system depends on the DVL information and the accuracy in linear motion is not as good as the SINS/DVL integrated navigation system. How to take advantage of both systems and make seamless changing between them are quite meaningful in real application field.

The organization of this paper proceeds as follows. The next section outlines the structure of AUV navigation systems. Section III details the error of normal Kalman filter in maneuvering state, after explaining the traditional adaptive Kalman filter, the dynamic sliding window is introduced to be used in the improved adaptive Kalman filter. Section IV proposes the maneuver detect algorithm, and the improved adaptive Kalman filter proposed in previous session will be used for SINS/DVL maneuvering state, and then introduces the design of whole structure of the navigation systems and the seamless changing method. In Section V, simulation is made to verify the design. Finally, conclusions are drawn.

II. THE STRUCTURE OF AUV NAVIGATION SYSTEM

The AUV navigation system is composed of two sub systems namely SINS/DVL integrated navigation system and Compass/DVL dead reckoning system. They will be discussed respectively as they are the foundation of the seamless changing method.

A. *SINS/DVL Integrated Navigation System*

The SINS in a stand-alone mode provides the position, velocity, and attitude information, thus the relative error information including the gyro drift are chosen as the state vector representing the characteristic of SINS in long-time working state. The gyro drift's error model is regarded as a first-order Markov process. The random time-varying drift caused by accelerometer is taken into account as the system noise. The corresponding linear error equation of SINS can be referred to[1]. Considering the error model of DVL, the state vector of Kalman filter is augmented with the DVL velocity offset's error δV_d, log misalignment angle error $\delta\Delta$ and scale coefficient error δC . δV_d and $\delta\Delta$ are expressed by first-order Markov process, δC is random constant drift.

The discrete model of the integrated SINS/DVL system is as follows:

$$\left.\begin{aligned} X_k &= \Phi_{k,k-1}X_{k-1} + \Gamma_{k,k-1}W_{k-1} \\ Z_k &= H_k X_k + v_k \end{aligned}\right\} \quad k \geq 1 \tag{1}$$

where X_k and W_k denote system's state vector and noise vector; Z_k and V_k denote system's measurement vector and noise vector; $\Phi_{k,k-1}$ and H_k denote state vector's transition matrix and measurement matrix; Γ_{k-1} denotes system's noise matrix.

The state vector and noise vector of SINS/DVL integrated navigation system are given by

$$X = \begin{bmatrix}\delta\varphi & \delta\lambda & \delta V_x & \delta V_y & \phi_x & \phi_y & \phi_z & \varepsilon_x & \varepsilon_y & \varepsilon_z & \delta V_d & \delta\Delta & \delta C\end{bmatrix}^T. \tag{2}$$

$$W = \begin{bmatrix}0 & 0 & a_x & a_y & 0 & 0 & 0 & w_x & w_y & w_z & w_d & w_\Delta & 0\end{bmatrix}^T. \tag{3}$$

Where $\delta\varphi, \delta\lambda$ denote latitude error and longitude error; $\delta V_x, \delta V_y$ denote east and north velocity error; V_x, V_y denote east and north velocity; ϕ_x, ϕ_y , ϕ_z denote north , east level and azimuth misalignment angle; ε_x , ε_y , ε_z denote gyro drift; a_x , a_y denote accelerometer random drift; $w_x, w_y, w_z, w_d, w_\Delta$ are stimulative white noise.

The difference between SINS velocity and DVL velocity compose the measurement vector.

$$Z = \begin{bmatrix}\delta V_x - \delta V_{dx} \\ \delta V_y - \delta V_{dy}\end{bmatrix} = HX + v \tag{4}$$

$$H = \begin{bmatrix}0 & 0 & 1 & 0 & 0 & 0 & -V_y & 0 & 0 & 0 & -\sin K_d & -V_y & -V_x \\ 0 & 0 & 0 & 1 & 0 & 0 & V_x & 0 & 0 & 0 & -\cos K_d & V_x & -V_y\end{bmatrix}. \tag{5}$$

B. *Compass/DVL Navigation System*

The Compass/DVL navigation system is somehow like a black box system, the dead reckoning working principle requires the outer sensors providing the initial position, and with the bearing information provided by compass and velocity information provided by DVL, the position will be calculated.

$$\varphi_{K(DR)} = \varphi_{K-1(DR)} + V_{dy}T / R_{yp} \tag{6}$$

$$\lambda_{K(DR)} = \lambda_{K-1(DR)} + V_{dx}T / (R_{xp}\cos(\varphi_{K(DR)})) \, . \tag{7}$$

Where $\varphi_{K(DR)}$ and $\lambda_{K(DR)}$ denote latitude and longitude information of dead reckoning system; V_{dx}, V_{dy} denote DVL east and north velocity; R_{xp} and R_{yp} denote the curvature radiuses of the reference ellipsoid in north-south and east-west directions.

Although the dead reckoning system is not as accurate as SINS/DVL system in long-term linear course, it's not sensitive to the vehicle maneuver. Later, we'll have a look at the seamless changing method to combine the two navigation system.

III. IMPROVED ADAPTIVE KALMAN FILTER

The estimation accuracy of traditional Kalman filter depends on priory knowledge about the system model and noise statistics. From the state equation, we could intuitively find out the relationship of the position, velocity and accelerate. With the help of transition matrix $\Phi_{k,k-1}$, the system could predict the state vector in next time according to the current state vector, which is the main cause of the Kalman filter error in the maneuvering state. Let's take a look at the case when the vehicle is going to change from curvilinear course to linear course. Theoretically, the predicted result is always based on the last state vector, so it is obvious the system would expect the vehicle turning a small arc again current position, while on the contrary, the real trajectory is the vehicle moving along the tangent direction of the last step. There's a disadvantageous factor that for the SINS/DVL integrated navigation system, there's no measurement like GPS signal to revise the position information, so the filtering result at the motion changing part is inevitably oscillating for a short period until the Kalman filter converging to its stable state. But for the close loop navigation system, it could be very dangerous that after a long-term working period, the SINS algorithm revised by the oscillating filtering information could cause the filter divergence at the motion changing state.

The solution for such problem is to minimize the effect of prediction part and maximize the effect of the innovation sequence in the maneuvering state. If we totally rely on the measurement of DVL velocity information, and drop the prediction part, then the calculation turns to the dead reckoning system. Then it comes to our next topic of seamless changing. However, for the SINS/DVL system, we enlarge the gain matrix according to the innovation sequence and adopt the improved adaptive Kalman filter for the maneuvering state shown in Fig. 1. Later, the traditional adaptive Kalman filter algorithm and improved adaptive Kalman filter will be discussed separately about the process noise unknown scenario.

A. Traditional adaptive Kalman filtering algorithm when Q is unknown

Let's take a look at the case of unknown precise information about covariance matrix of process noise covariance $Q = M[w_k w_k^T]$ is unknown, after rewriting the Kalman filter,

$$\hat{x}_k = \Phi_{k,k-1}\hat{x}_{k-1} + \Gamma_{k-1}\hat{w}_{k-1} \tag{8}$$

$$\hat{x}_k - \Phi_{k,k-1}\hat{x}_{k-1} = \Gamma_{k-1}\hat{w}_{k-1} \tag{9}$$

$$\hat{x}_k - \Phi_{k,k-1}\hat{x}_{k-1} = K_k v_k \tag{10}$$

$$\Gamma_{k-1}\hat{w}_{k-1} = K_k v_k \tag{11}$$

$$\Gamma_{k-1}M[\hat{w}_{k-1}\hat{w}_{k-1}^T]\Gamma_{k-1}{}^T = K_k M[v_k v_k^T]K_k^T \quad . \tag{12}$$

Where $v_k = Z_k - H\hat{X}_{k,k-1}$ denotes innovation sequence, and K_k denotes the gain matrix. Using the formula for the Gaussian probability density, the covariance matrix of innovation sequence can be defined as [5]:

$$\hat{C}_k = M[\hat{w}_{k-1}\hat{w}_{k-1}^T] = v_k v_k^T \quad . \tag{13}$$

$$\Gamma_{k-1}\hat{Q}_{k-1}\Gamma_{k-1}^T = K_{k-1}\hat{C}_k K_{k-1}^T \tag{14}$$

B. Improved Adaptive Kalman fitering algorithem when Q is unknown

Many researchers engage in the study of the adaptive Kalman filter. The covariance scaling algorithm was used to improve the stochastic modeling [7]. MMAE (Multiple Model Adaptive Estimation) algorithm using multiple Kalman filters running simultaneously to solve the uncertainty of the modeling problem was discussed [8] [9]. Fuzzy logic was used to tune the Kalman filter in the integrated navigation system [10]. Many algorithms basing on the IAE (Innovation Adaptive Estimator) were investigated [6] [11] [12] [14]. MMAE hasn't be widely used because of its complicated calculation. To estimate the process noise on-line, this paper proposes a new method basing on the innovation adaptive estimator: the sliding window length is adjusted by the measurement remnant value automatically, thus the measurement covariance could be tuned according to innovation message.

For a stationary system, one can determine the following estimate of C_k [5] [6]:

$$\hat{C}_k = \frac{1}{k}\sum_{i=1}^{k} v_i v_i^T \tag{15}$$

or in recurrent

$$\hat{C}_k = \frac{k-1}{k}\hat{C}_{k-1} + \frac{1}{k}v_i v_i^T \tag{16}$$

From (15)-(16), we could see that the traditional way to estimate C_k is the mean value of innovation sequence from the beginning to the current time K. Then with the time passing by, the effect of latest innovation method become smaller and smaller $\frac{1}{k}v_i v_i^T$, if the vehicle motion changing sharply at that time, the innovation sequence almost has no effect on the measurement covariance estimator.

Fading memory algorithm has been introduced to the AKF by some researchers, and the fixed length N (filtering window) to calculate covariance matrix of the innovation sequence has been used to increase the effect of the measurement value [12].

$$\hat{C}_k = \frac{1}{N}\sum_{i=k-N}^{k} v_i v_i^T \tag{17}$$

We could imagine that when the process noise has a sharp change due to vehicle maneuver, we would like to enhance the effect of the innovation sequence to evaluate the process noise online; while on the contrary, we would like to keep the stationary characteristic of the system.

After investigating the disadvantage of both Saga-Husa and fading memory algorithm, this paper proposes a method to adjust the window length automatically.

$$d = v_k^T C_k^{-1} v_k = (Z_k - H\hat{X}_{k,k-1})^T (H_k P_{k+1/k} H_k^T + R_k)^{-1} (Z_k - H\hat{X}_{k,k-1}) \tag{18}$$

Then the measurement remnant value d is used to adjust the window length which could be called "tuning factor". The sliding window length becomes smaller as the remnant value goes bigger, obviously, the minimum length of the sliding window is 1, and the maximum length of it is k, and the innovation sequence takes the biggest effect and smallest effect for above scenarios.

$$N = 1 \,, \quad d > \alpha_{\max}$$

$$N = k \,, \quad d < \alpha_{\min}$$

$$N = Integer(k \times \lambda^{d-\alpha_{\min}}) \,, \lambda < 1, N \geq 1 \,, \quad \alpha_{\min} < d < \alpha_{\max} \qquad (19)$$

Where, α is the given threshold, the value of λ decides the convergence speed of the length N by the remnant value. The sliding window length could be adjusted automatically according to (19). After substituting sliding window length N to (18) and the process noise could be calculated according to (14).

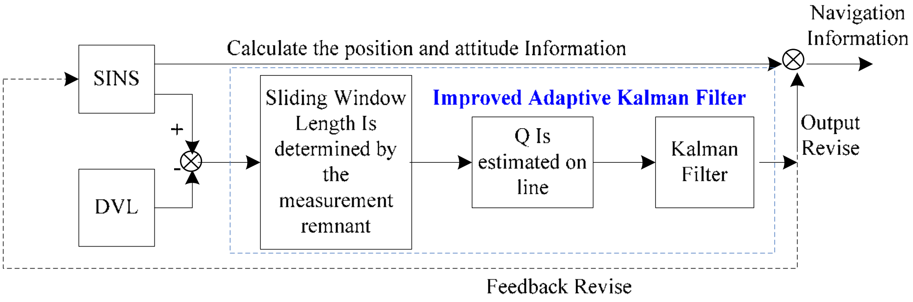

Navigation Information
Calculate the position and attitude Information
SINS
+
-
DVL
Sliding Window Length Is determined by the measurement remnant
Improved Adaptive Kalman Filter
Q Is estimated on line
Kalman Filter
Output Revise
Feedback Revise

Figure 1. SINS/DVL Simple Structure with Improved Adaptive Kalman Filter

The improved adaptive Kalman filtering algorithm of SINS/DVL navigation system then comes as Fig.1.

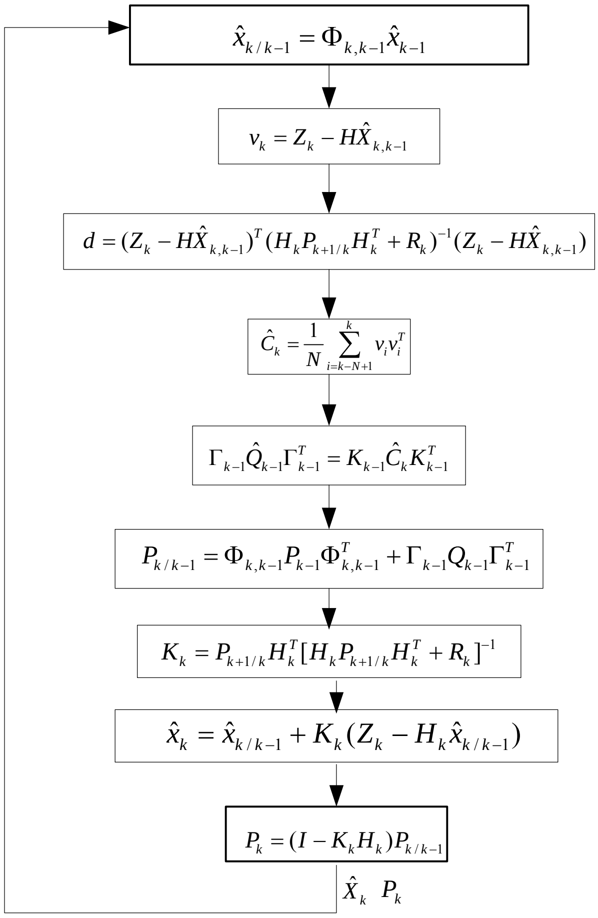

$\hat{x}_{k/k-1} = \Phi_{k,k-1}\hat{x}_{k-1}$
$v_k = Z_k - H\hat{X}_{k,k-1}$
$d = (Z_k - H\hat{X}_{k,k-1})^T (H_k P_{k+1/k} H_k^T + R_k)^{-1} (Z_k - H\hat{X}_{k,k-1})$
$\hat{C}_k = \frac{1}{N} \sum_{i=k-N+1}^{k} v_i v_i^T$
$\Gamma_{k-1}\hat{Q}_{k-1}\Gamma_{k-1}^T = K_{k-1}\hat{C}_k K_{k-1}^T$
$P_{k/k-1} = \Phi_{k,k-1}P_{k-1}\Phi_{k,k-1}^T + \Gamma_{k-1}Q_{k-1}\Gamma_{k-1}^T$
$K_k = P_{k+1/k}H_k^T[H_k P_{k+1/k} H_k^T + R_k]^{-1}$
$\hat{x}_k = \hat{x}_{k/k-1} + K_k(Z_k - H_k\hat{x}_{k/k-1})$
$P_k = (I - K_k H_k)P_{k/k-1}$
$\hat{X}_k \; P_k$

Figure 2. Improved Adaptive Kalman Filter Algorithm with Sliding Window

The advantage of above method is that the ratio of the prior innovation message depends on the latest measurement remnant value, the current innovation sequence could work effectively in this way, and when the remnant value is small enough, the sliding window length is extended to guarantee the stationary characteristic of the system. Thus the aim of adjusting the sliding window length is realized. The sliding window idea has been raised by author to estimate the measurement noise covariance on line [14]. However, there's still something to be done to separate the situation whether the process noise or the measurement noise is changed by some methods, e.g., judge the consistency of sensor information before it is used. Further research will be done in future. In next session, the improved adaptive Kalman filter algorithm will be used when maneuver event is detected to realize the smooth transition of AUV navigation.

IV. DESIGN OF SEAMLESS CHANGING METHOD

A. Maneuver Detection

Many discussions about maneuver detection have been made for Strong Target Tracking scenario in literature. The most commonly used method is the measure remnant χ^2 method [13]. However, this method could be disturbed by a fault instead of maneuver.

Let's take $D_k = H_k P_{k+1/k} H_k^T + R_k$, then we could rewrite (18)

$$d = v_k^T D_k^{-1} v_k \qquad (20)$$

Compare it with the given threshold α, if

$$d > \alpha \qquad (21)$$

The traditional method will assume the vehicle motion state as maneuver.

To avoid the misjudgment, the consistent detection of bearing information is used as supplement for maneuver detection. The heading sensor could record the bearing information of the last three steps: $\phi_{zK-2(DR)}$, $\phi_{zK-1(DR)}$, $\phi_{zK(DR)}$, considering the wavelet effect, we could assume that

$$\left|\phi_{zK-2(DR)} - \phi_{zK-1(DR)}\right| \approx \left|\phi_{zK(DR)} - \phi_{zK-1(DR)}\right| \approx \omega_{yaw}T \quad . (22)$$

Where ω_{yaw} is the yaw angular velocity against up axis caused by wave stimulation. When the first detection by (21) is given, and (22) is satisfied, we could assume the vehicle is in maneuver state. Otherwise, the (21) is caused by DVL fault, as the inertial sensor is normally taken to be reliable, and then the SINS/DVL should use prediction mode.

B. the Seamless Changing Method

Let's assume the AUV starts with linear course, then SINS/DVL is used to output navigation information.

Fig. 3 could help to understand the thorough design of AUV navigation systems.

SINS calculates position, velocity and attitude according to the output of the accelerator and gyro output. In order to realize the optimal filtering algorithm, when the AUV moves in straight line, the conversional Kalman Filter will be adopted for the SINS/DVL integrated navigation systems, and the estimated information will be used to revise relative information of SINS. When the maneuvering state is judged by the previous "maneuver detect" algorithm, the sliding window length will be determined by the measurement remnant, and the improved adaptive Kalman filtering algorithm will be used by the SINS/DVL integrated navigation systems. Please note that when maneuver is detected, the SINS/DVL uses adaptive Kalman filter, the last position of SINS/DVL is used to reset the compass/DVL dead reckoning system, and the DR system is adopted as the AUV navigation output resources which will be used as the feedback information to revise SINS.

When maneuver judgment tells us the vehicle is changing back to straight line, AUV keeps using DR (dead reckoning) system for several steps, after that, AUV adopts SINS/DVL as the output navigation system, and the SINS/DVL switches back to normal Kalman filtering process. The structure of AUV navigation system could be referred to Fig. 3 and Fig. 4.

The advantage of this design is that there's no extra cost on expensive high-level accuracy sensors, and the positioning accuracy will be improved in maneuvering state just with normal sensors. And the navigation method of SINS/DVL integrated navigation system combines the optimal Kalman filter in straight line and the dynamic adaptive Kalman filter in maneuver state fairly well, which could reduce the oscillation when switching between different modes.

V. COMPUTER SIMULATION

We now describe the application of the seamless changing method for navigation of Autonomous Underwater vehicles. In order to investigate the performance of the new approach, the medium precision's measurement sensors are simulated. The measurement sensors' errors and the initial navigation parameter's errors are given as (21). After fine alignment, the misalignment angle in three axes (pitch, roll, yaw error) is given as 0.0005°, 0.0005°, and 0.003° separately. The initial quaternion is calculated according to the misalignment angle so that the vehicle body frame is aligned with the local geographical frame. The gyro drift is chosen as $1\times10^{-3^\circ}/h$. The accelerometer parameters are given by 10^{-4}g. The velocity offset's error δV_d and log misalignment angle error caused by the ocean current $\delta\Delta$ can be chosen as

$$\beta_i^{-1} = 2\text{h}(i = x, y, z), \beta_d^{-1} = 5\min; \beta_\Delta^{-1} = 15\min$$

The whole system takes both advantage of SINS/DVL integrated navigation systems and the DR navigation systems. Thus, the AUV maneuver capability is improved with such seamless changing.

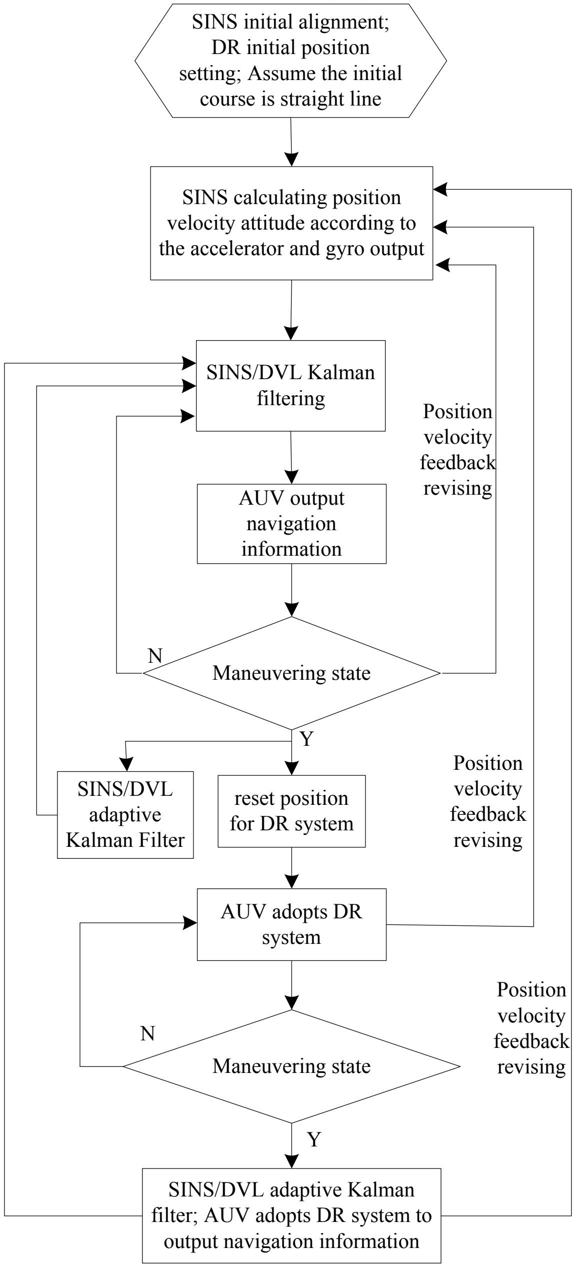

SINS initial alignment; DR initial position setting; Assume the initial course is straight line
SINS calculating position velocity attitude according to the accelerator and gyro output
SINS/DVL Kalman filtering
Position velocity feedback revising
AUV output navigation information
N
Maneuvering state
Y
SINS/DVL adaptive Kalman Filter
reset position for DR system
Position velocity feedback revising
AUV adopts DR system
Position velocity feedback revising
N
Maneuvering state
Y
SINS/DVL adaptive Kalman filter; AUV adopts DR system to output navigation information

Figure 3. Diagram of the whole structure

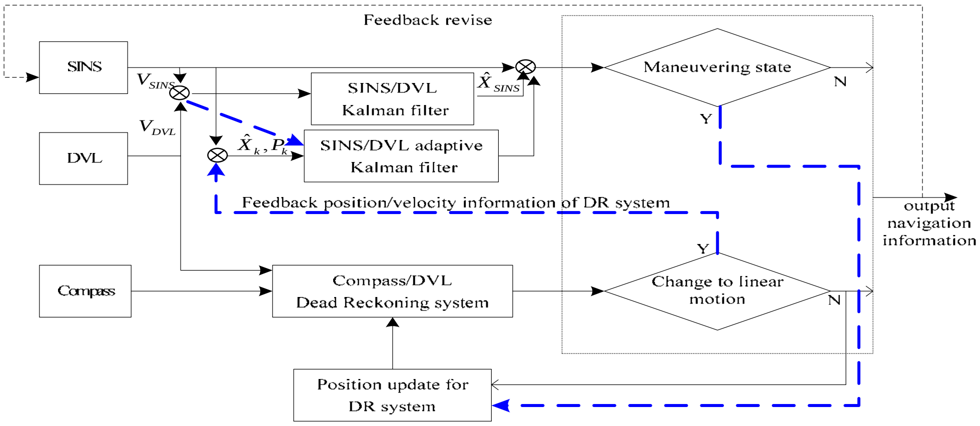

Feedback revise
SINS
DVL
Compass
SINS/DVL Kalman filter
SINS/DVL adaptive Kalman filter
Compass/DVL Dead Reckoning system
Position update for DR system
Maneuvering state
Change to linear motion
Feedback position/velocity information of DR system
output navigation information
Y
N
Y
N

Figure 4. Structure of AUV navigation system

The vehicle trajectory is composed of several courses: straight forward with the bearing angle of 10°, velocity 10knots, lasting for 10 minutes; maneuver state, turning around a circle with 500m radius, at 10knots rate, 3.5 π arc; changing back to straight line, lasting for 10 minutes, velocity 10knots.

The sway motion model is represented as below:

$$\left.\begin{array}{l}\theta=\theta_m\sin(\omega_\theta t+\theta_0)\\ r=r_m\sin(\omega_r t+r_0)\\ \varphi=\varphi_m\sin(\omega_\varphi t+\varphi_0)\\ V_x=V_{x0},\quad V_y=V_{y0}\end{array}\right\} \tag{23}$$

where θ, r, φ are pitch, roll and yaw respectively; the sway magnitudes θ_m, r_m, φ_m are 5°, 6°, and 5° respectively. The sway period of pitch, roll and yaw is 10s, 10s, and 10s respectively. The initial phase angles θ_0, r_0 and φ_0 are set to 0°.

The trajectory of AUV is shown in Fig.5 and zoom in figure of the first motion changing trajectory is shown in Fig. 6. From which, we could see that the output of AUV navigation system is quite smooth without big gap during the motion change. The gyro output of three axes is shown in Fig.6.

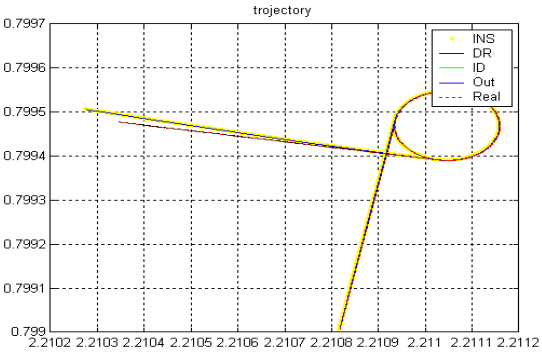


Figure 5. trajectory of AUV

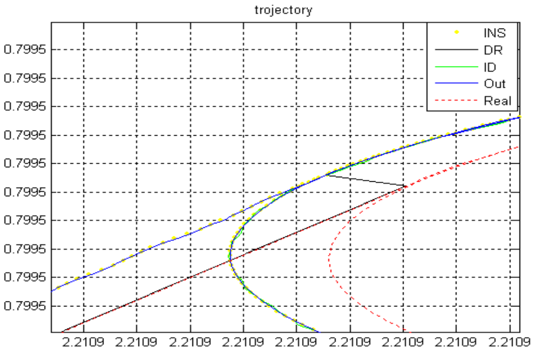


Figure 6. zoom in figure of the first motion change trajectory

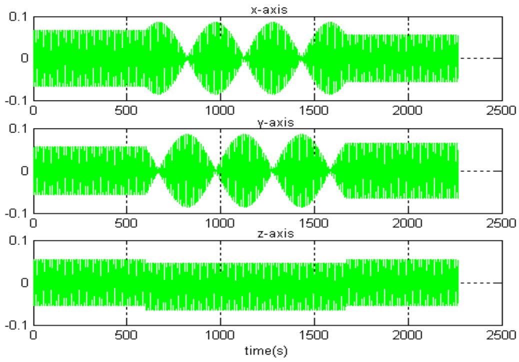


Figure 7. Gyro ouput of three axes

Fig.8 and Fig.9 reveal the longitude estimation error; Fig.10 and Fig.11 reveal the latitude estimation error. From these four figures, we could intuitively find out that the oscillation by the adaptive SINS/DVL Kalman filter is not as sharp as normal Kalman filter at the motion changing state, and the magnitude goes smaller during the maneuver period, while on the contrary, the general position output information by the method given in this paper is quite smooth due to the Compass/DVL dead reckoning system has a better performance in the maneuver state. Fig.12 and Fig.13 reveal the east and north velocity error of the system.

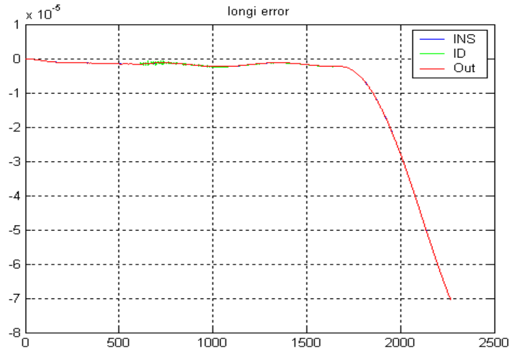


Figure 8. longitude error for maneuvering state

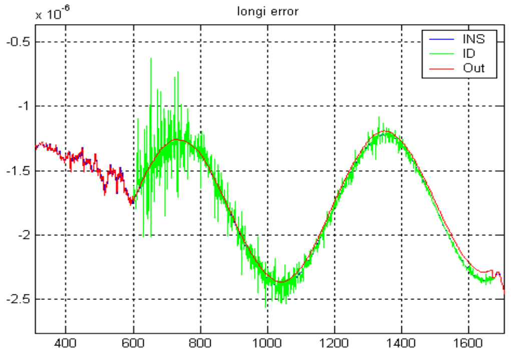


Figure 9. zoom in plot of longitude error

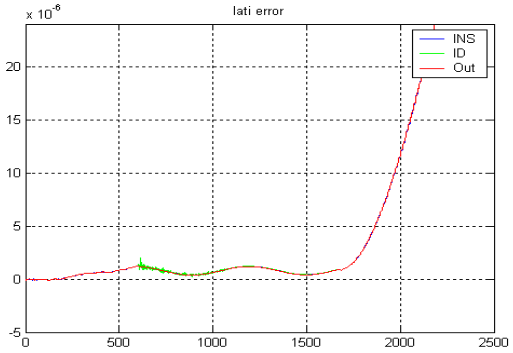


Figure 10. latitude error for maneuvering state

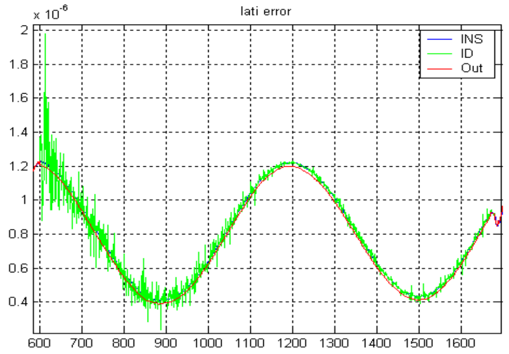


Figure 11. zoom in plot of latitude error

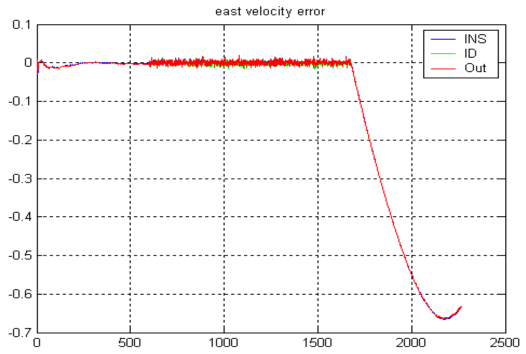


Figure 12. east velocity error

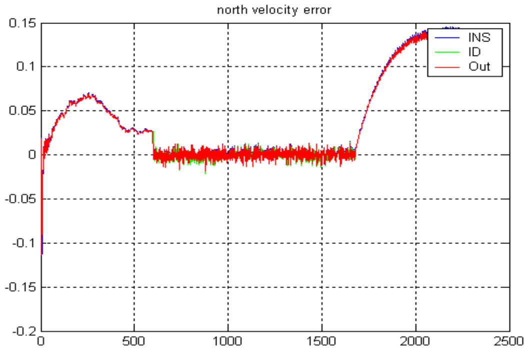


Figure 13. north velocity error

VI. Conclusion

As AUV faces a serious problem of maneuvering state in marine environment, this paper proposes a new method to improve the maneuver capability of the AUV integrated navigation systems. The main target is to reduce the oscillation of traditional Kalman filter in motion changing state and improve the positioning accuracy. The whole design is based on the normal sensor without extra cost.

It is founded that the bearing consistent check combining with the measure remnant χ^2 method could avoid the misjudgment of maneuver. The dynamic sliding window tuned by measurement remnant is proposed here to judge the process noise on line, and thus is the main idea of the improved adaptive Kalman filter which could help the SINS/DVL system to reduce oscillation when the AUV is in motion changing state. The whole structure of the AUV navigation systems is introduced with a thorough diagram. The most important of all is that after adopting the seamless changing method, the AUV output navigation information is much smoother than before. The simulation result proves that such design with improved adaptive Kalman filter is efficient and could be used in the application field.

Acknowledgment

The financial support from the National Natural Science Foundation of China under the grant No. 60674087 is gratefully acknowledged.

REFERENCES

[1] L. Stutters and H. Liu, “Navigation Technologies for Autonomous Underwater Vehicles”, *IEEE Trans. on Systems, man and cybernetics, Part C: Applications and Reviews*, vol. 38, pp. 581-589, July 2008.

[2] S. Ishibashi, S. Tsukioka, H. Yoshida, T. Hyakudome, T. Sawa, J. Tahara, T. Aoki and A. Ishikawa, “Accuracy Improvement of an Inertial Navigation System Brought about by the Rotational Motion”, *Japan Agency for Marine-Earth Sci. & Technol. (JAMSTEC), Yokosuka*, pp. 1-5, June 2007

[3] B. Liu and J. Fang, “A New Adaptive Feedback Kalman Filter Based on Observability Analysis for SINS/GPS”, *Acta Aeronautica et Astronautica Sinica*, vol. 29(2), 2008.

[4] Z. Guo, F. Sun, “Research on Integrated Navigation Method for AUV”, *Journal of Marine Science and Application*, vol.4, no.2 , pp. 34-38, June 2005

[5] Oleg S. Salychev, Applied Inertial Navigation: Problems and Solutions, Bauman MSTU Press, .Moscow, 2004

[6] M. Bai, X. Zhao and Z. Hou, “Application of an Adaptive Filter in Initial Alignment of Strapdown Inertial Navigation System with Large Misalignment Error”, *Chinese Journal of Sensors and Actuators*, vol. 21, no. 6, pp. 1066-1069, Jun 2008.

[7] C. Hu, Y. Chen, and W. Chen, “Adaptive Kalman filtering for DGPS positioning”, in *Proceedings of the International Symposium on Kinematic Systems in Geodesy*, Geometrics and Navigation (KIS), Canada, 2001

[8] R.B.Brown and P.Y. C. Hwang, “Introduction to Random Signals and Applied Kalman Filtering”, 3rd Ed. John Wiley and Son Inc. 1997

[9] K. A. Fisher and P. S. Maybeck, “Multiple Model Adaptive Estimation with Filter Spawning” [J], *IEEE Trans. Aerospace and Electronic Systems*, vol. 38, no. 3, pp. 755-768, 2002.

[10] A. Hiliuta, R. JR. Landry, and F. Gagnon. “Fuzzy Corrections in a GPS/INS Hybrid Navigation System”, *IEEE Transactions on Aerospace and Electronic Systems*, vol. 40, no. 2, pp. 591-600, April 2004

[11] C. Hu, W. Chen, Y. Chen and D. Liu, “Adaptive Kalman Filtering for Vehicle Navigation”, *Journal of Global Positioning Systems*, vol. 2, no. 1, pp. 42-47, 2003.

[12] X. Chen, D. Yang, K. Zhai, “Research on the AEKF Applied on Attitude Determination”, *Aerospace Control*, vol. 26, no. 1, pp. 47-51, Feb.2008.

[13] Q. Wang, Qing, W. Liu and C. Dong, “An H ∞ Suboptimal Filter Target Tracking Algorithm W ith Acceleration Compensation,” *Journal of System Simulation*,vol.18, no.7, pp.2042-2045, July.2006.

[14] Y. Hao, Z. Guo, F. Sun and W. Gao, “Adaptive Extended Kalman Filtering for SINS/GPS Integrated Navigation Systems”, unpublished.

Zhen Guo, born in 1980, received her master degree in 2005 and just working for her P.H.D scholarship in college of Automation, Harbin Engineering University. The main research directions are navigation guidance and control.

Yanling Hao, born in 1944. She’s the professor and PhD Candidate Supervisor of HEU automation college. She has been engaged in the research work of navigation guidance and control. She has led many projects in many research fields which are aided by national natural science foundation of China.

Feng Sun, born in 1944, was graduated from precise instrument major in Harbin Institute technology in 1967. He’s the professor and PhD Candidate Supervisor of HEU. He has been engaged in the research work of instrument and machine for many years, and he has also taken charge of mange scientific research projects.

A Hybrid Password Authentication Scheme Based on Shape and Text

Ziran Zheng
School of Management & Economics
Shandong Normal University, Jinan, China
Email: zzrnature@gmail.com

Xiyu Liu
School of Management & Economics
Shandong Normal University, Jinan, China
Email: zzrnature@gmail.com

Lizi Yin
School of Science
University of Jinan, Jinan, China
Email: ss_yinlz@ujn.edu.cn

Zhaocheng Liu
Department of Management
Jinan Railway Polytechnic, Jinan, China
Email: liuzhch100@163.com

Abstract—**Textual-based password authentication scheme tends to be more vulnerable to attacks such as shoulder-surfing and hidden camera. To overcome the vulnerabilities of traditional methods, visual or graphical password schemes have been developed as possible alternative solutions to text-based password schemes. Because simply adopting graphical password authentication also has some drawbacks, schemes using graphic and text have been developed. In this paper, we propose a hybrid password authentication scheme based on shape and text. It uses shapes of strokes on the grid as the origin passwords and allows users to login with text passwords via traditional input devices. The method provides strong resistant to hidden-camera and shoulder-surfing. Moreover, the scheme has high scalability and flexibility to enhance the authentication process security. The analysis of the security level of this approach is also discussed.**

Index Terms—**first term, second term, third term, fourth term, fifth term, sixth term**

I. Introduction

How to increase the level of authentication security has become an important problem in the age of information. The most general authentication methods in computers and other devices require the submissions of the users' names and their passwords. The most serious problem about textual password is the vulnerabilities to various attacks. Due to the fact that this type of scheme is based on the characters, the login passwords are quite easy to guess and if the passwords get longer, they become harder to remember for the users themselves. To overcome the vulnerabilities of textual passwords, visual or graphical password schemes have been developed as possible alternative methods to the traditional authentication process. The main idea of graphical passwords is to use the images or shapes to replace the text, since graphical signs are easier to remember than pure characters [1].

Although graphical password schemes have been considered as alternatives to traditional text password, they also have some drawbacks. For example, some of them have vulnerabilities to shoulder-surfing because of the users' direct actions upon the input screen. And some schemes will require users to input the password for several times. In addition, most of the graphical schemes have far more complexity in the implementation of the application.

This work is proposed to make a bridge between the graphic and text password. Since the shape as the password have larger space and easier to remember, we take the advantage of the shape as the users' original passwords. To make the implementation easy and avoid direct interaction appeared between the user and screen, a grid with characters is adopted to construct the new system. With this new authentication scheme, users can only just remember the shapes and strokes they like as their passwords. However, the system authenticates the shape passwords just with text on the grid and their input order during the process.

What we focus on the design of the scheme are as follows: (1) Using shapes and strokes on the grid as the original password, since the shape of stroke can be easier

doi:10.4304/jcp.5.5.765-772

to remember than text. (2) Text-based login process, which supports keyboard as the input device. (3) Strong resistant to shoulder-surfing and hidden camera. (4) It has large password space and robust mechanism against the brute force attack.

This paper is organized as follows. In the next section the background and related work about the graphical scheme are introduced. In the section after that the basic and further description of our approach is presented. The next section discusses and analyzes the security level of the scheme. The conclusion and future work are presented in the last section.

II. Back Ground and Related Work

A graphical password scheme, in which a password is generated through asking the user to click on a graphic or an image provided by the system, is designed by Blonder [2]. When creating a password, the user is asked to choose four images of human faces from a face database as their own password. In the authentication stage, users must click on the approximate areas of those locations. This method is considered as a more convenient password scheme than textual scheme, for the image can help users to recall their own passwords. Wiedenbeck, et al. [3] extended the approach and proposed a system called "PassPoint". It allows users to click on any locations on the image to create the passwords. The system will calculate a tolerance around each pixel which has been chosen. The users must click within the tolerance of the chosen pixels.

Jansen [4-6] proposed a graphical password scheme for mobile devices. During enrollment, a user is asked to choose the theme consists of photos in thumbnail size and set a sequence of pictures as a password. In the authentication stage, a user must input the registered images in the correct order. Each thumbnail image is assigned a numerical value, thus the sequence of the chosen ones will create a numerical password. Because the number of picture is limited to 30, the password space of this scheme is not large.

Jermyn, et al [7] proposed a technique call "Draw a Secret (DAS)". This system allows users to create their own passwords by drawing something on a 2D grid. When a user finishes the drawing, the system stores the coordinates of the grids occupied by the picture. During authentication, users must re-draw the picture which had been created by them. The user will be authenticated if the drawing touches the same gird in the right order. The password space of this scheme is proved to be larger than the full text-based password space.

Thorpe and van Oorschot [8] analyzed the memorable password space of the DAS. Graphical dictionaries were introduced and possibilities of a brute-force attack using dictionaries are studied. They showed that a significant fraction of users will choose mirror symmetric password, since people recall symmetric images better than asymmetric images. Thorpe and van Oorschot [9] also studied the impact of password length and stroke-count as a complexity property of the DAS scheme. In order to improve the security, a "Grid Selection" technique is proposed. It allows users to select a rectangle region as the drawing grid, in which they may input the password. This method increases the DAS password space significantly. Further research was studied by Nali and Thorpe [10].

To overcome the shoulder-surfing problem, many techniques were proposed. Zhao and Li [11] proposed a shoulder-surfing resistant scheme "S3PAS" The main idea of the scheme is as follows. In the login stage, they must find their original text passwords in the login image and click inside the invisible triangle region. The system integrates both graphical and textual password scheme and has high level security. Man, et al, [12] proposed another shoulder-surfing resistant technique. In this scheme, a user chooses many images as the pass-objects. The pass-objects have variants and each of them is assigned to a unique code. In the authentication stage, the user must type the unique codes of the pass-objects variants in the scenes provided by the system. Although the scheme shows perfect results in resisting hidden camera, it requires the user to remember code with the pass-object variants. Further research based on this method was conducted in [13].

Luca, et al. [14] proposed a stroke based shape password for ATMs. They argued that using shapes will allow more complex and more secure authentication with a lower cognition load.

More graphical password schemes have been summarized in a recent survey paper [15].

III. Hybrid Password Scheme

The hybrid password scheme based on shape and text is designed not only for the traditional computers but can be used in the mobile devices. The basic idea of our scheme is to make a map from shape to text with strokes of the shape and a grid with text. The map could be constructed quite simple and straight-forward. This mapping not only guides the user to master this scheme with ease, but makes the whole system easy to implement.

Fig. 1 shows the idea of this work. Users should just think some personal shapes and its strokes as their origin password and enter character in the authentication as the login password.

The whole process includes two main steps: the password creation step, and the login step. In the basic scheme, we take a simple example to descript the two stages. Variants of the scheme will be introduced in the further description.

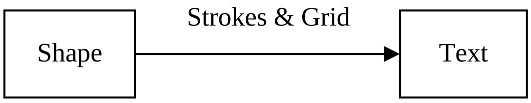


Figure 1. Mapping from shape to text through strokes and grid

A. *Notations of the Scheme*

The following notations, which are used throughout the paper, are defined to help the presentation and analysis of the scheme.

- *U*: The set of elements appeared in the grid in the interface.
- *V*: Input passwords vector, which consists of elements in U.
- |*V*|: Size of the *V*. It also represents the length of the input passwords, or the strokes' size.
- *g*: the size of the grid.
- *S*: Shape of the password. For example, it could be "N", "1", "&"or any other forms. *S** means the number of different types of the shape S.
- |*S*|: Number of strokes of the password.
- *H*: The password space. H_t and H_s represent the text-based and stroke-based password space respectively.

B. Basic Scheme

In the first step, the user is asked to select a group of elements on the grid shown in the interface as the original password. In this example, we use $g = 5\times 5$ gird to show the process. The password-set interface is shown in Fig. 2.

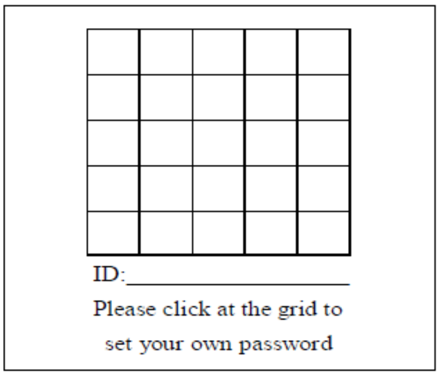

ID:_________________
Please click at the grid to
set your own password

Figure 2. Password set interface

Note that the size or the grid (*g*) can be different to meet the certain requirements and it could affect the security level of the scheme. More descriptions about this will be explained in the next section.

Firstly, a user is proposed to pick a shape S such as a number shape, a geometric shape, a character shape or even a random shape as his(or her) own original password. The criterion of choosing the shape is as easy to remember as possible for the users themselves. Though the number of the shapes could vary, it is not the key factor to the scheme. Thus we use one shape to describe this instance for the sake of convenience.

After the password shape is selected in their mind, the user should click on the grid in the interface following the shapes' stroke sequence. The system will store the shape and the order with the grid as the user's mapped text password.

Note that, this process doesn't have the same level of security than the login step, since the direct action between the user and the screen. And if the input device is the keyboard like the ATM, the password set process will reflect the original password of the user, if this process is recorded by the camera, the whole password and this scheme will not work. However, this disadvantage can be overcome by a multi-set process, which will be described in detail in next section.

Go back to the example, the user John chooses one of characters of his name "N" as the shape of the password. Suppose the sequence of the stroke "N" is in a simpler order than normal as the shape's stroke order.

When the shape and the order setting are finished, John could design the stroke on the grid as he likes (this is a mechanism to level up the security level. Even if the shape is known by the hacker in some way, the hacker would not be sure the shape's shape on the grid specifically). Here, we suppose that the shape is drawn fully at the grid. After that, the user clicks on the grid to form "N" as the original password. The set procedure can be seen more clearly in the Fig. 3.

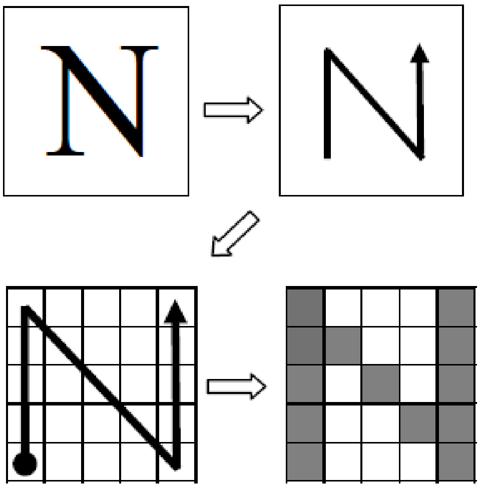
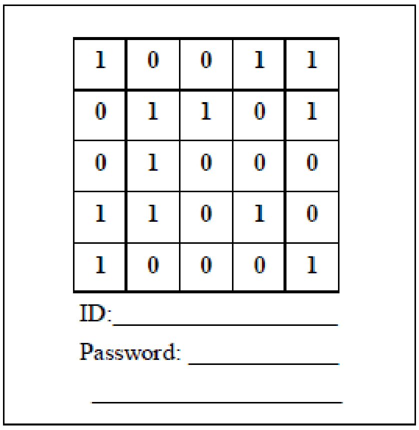


Figure 3. Password set procedure

Fig. 3 not only shows the procedure of the setting password, but also provides the idea of mapping from a simple shape into a grid. The shape is finally represented by a number of blocks on the grid.

In the login step, the interface is presented with a different style. The grid is filled with some similar symbols such as some numbers or characters. The feature of the approach here is to use quite a few numbers of the symbols, which consists of U. Since the less we use, the faster and more secure of the authentication process will be. Here we use the number "0" and "1" to show the example, which means *U* = {0, 1}. Note that the system will choose the symbol randomly from U to fill every grid. The login interface is shown in Fig. 4.



1	0	0	1	1
0	1	1	0	1
0	1	0	0	0
1	1	0	1	0
1	0	0	0	1

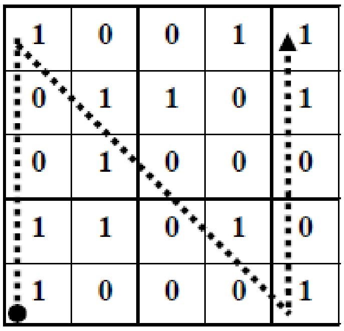
ID:__________________

Password: _____________

Figure 4. Login interface

During the authentication stage, the user John was asked to enter the password. He will use the keyboard with only "0"and"1" keys to input the password. The order and content of the password is entering the number in the grid following the original password shape's strokes which he has chosen in the password-set step.

Fig. 5 shows the image appeared in the John's mind, which is not the action or the image in the authentication scheme. It just helps to understand what the users would recall and think in the login step.



1	0	0	1	1
0	1	1	0	1
0	1	0	0	0
1	1	0	1	0
1	0	0	0	1

Figure 5. Original stroke on the interface

While looking at the number filled in the grid of the original shape, John should enter numbers in the right order. Thus, the password is as follows: 1100110110011, where V=[1,1,0,0,1,1,0,1,1,0,0,1,1].

The system will check if the input vector matches the numbers appeared in John's original sequence of the grid upon the interface created by the system. Because the texts with which the user enters are only using two keys, the login process is quite convenient. It is very useful to shorten the login process. More importantly, the act of inputting with only two keys can effectively resistant to the shoulder surfing.

If the password entered is not correct, then the system will generate another login interface grid for the user with characters randomly selected again. The symbols from U appeared in the grid varies at each login step, which means that the shape and the sequence of shape will not vary but the mapped text will not be the same at different interfaces. It also means the text passwords the user will input are not the same one at different login times. If hackers record the text the user input exclusively, they would get nothing about the information of any user's original password. Thus the text-based brute force attack with the "1"s and"0"s are useless.

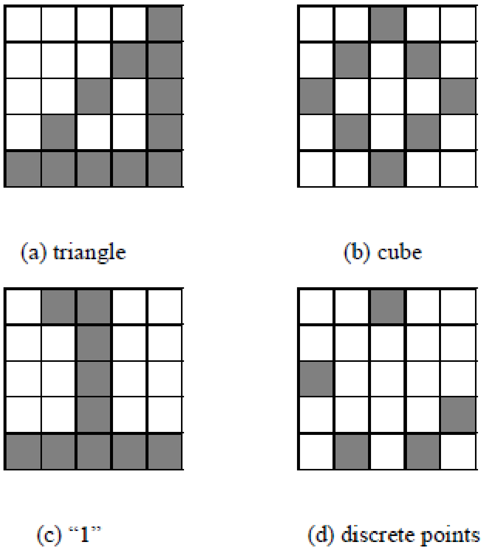
The main idea of the scheme is making the stroke shape as the password using the textual input. And we use this mechanism to resist the spy attack. The basic scheme is quite simple. To enhance the scheme, there are some points to explain in details.

IV. Further Descrition

The basic scheme can be extends from different respects. Some of them are made the system easy to use for the users and some are designed to improve the security of the whole system.

A.Shapes Choosing

At the set step, the shape S as the original password can be of quite different types. Users not only can choose the character but can also adopt the geometric shapes, the number shapes, the symbol shapes and even the arbitrary shapes as the preferred password shape. Fig. 6 shows several different kinds of shape: triangle, cube, number"1" and some discrete plots. This mechanism is to offer various alternatives for users.


(a) triangle (b) cube

(c) "1" (d) discrete points

Figure 6. different original shapes

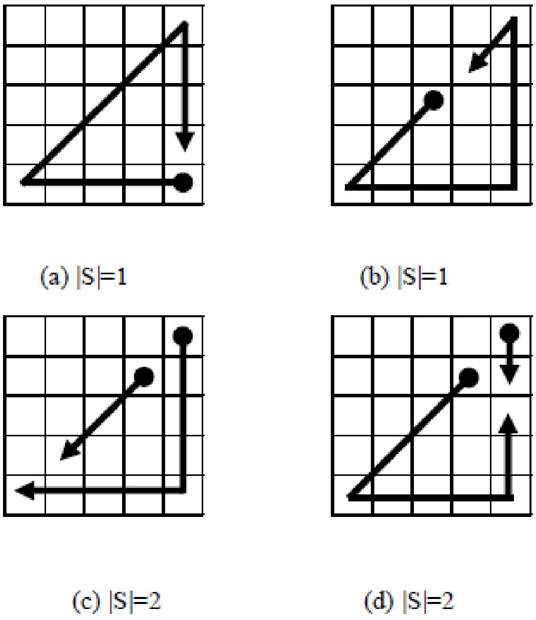
One shape also can have different styles, which means a conceptual shape will generate various specific styles. This description will be provided in the security analysis section.

Essentially, the shape that appeared in the user's mind is the blocks on the grid for the system. Theoretically, any boxes in the grid can be adopted as the user's original shape and the shape or shapes can have no meaning at all except for the user himself. Any unique shape with personality is considered as the better option than the normal shape with meanings.

B.Shokes Choosing

Even for the same shape, the password shape space is very large since it considers the sequence of the stroke and the number of the strokes |*S*| during the shape creating step.

Take the triangle shape for example. The stroke of the shape could have several variants, which are shown in Fig. 7. Black point means the start of the stroke and the arrow means the end. The shape triangle can be divided into several strokes. Also, these four figures are not all the variants.



(a) |S|=1

(b) |S|=1

(c) |S|=2

(d) |S|=2

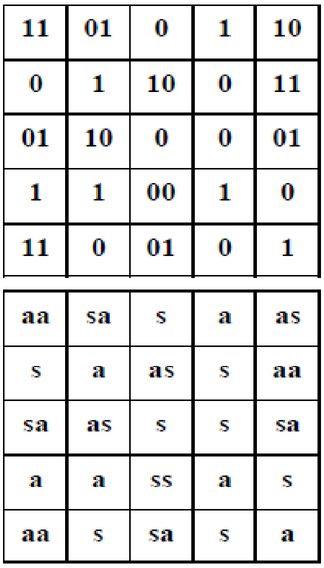
Figure 7. Stroke variants of triangle

Just like the shapes choosing, users could select any order and number of the strokes of their original shape password.

C.Different Interface

At the login step, the characters appeared in the grid can be in any form preferred by the designer or the user. In addition, the number in each grid could vary at the same time.

For example, Fig. 8 shows the variants of the login form.



11	01	0	1	10
0	1	10	0	11
01	10	0	0	01
1	1	00	1	0
11	0	01	0	1

aa	sa	s	a	as
s	a	as	s	aa
sa	as	s	s	sa
a	a	ss	a	s
aa	s	sa	s	a

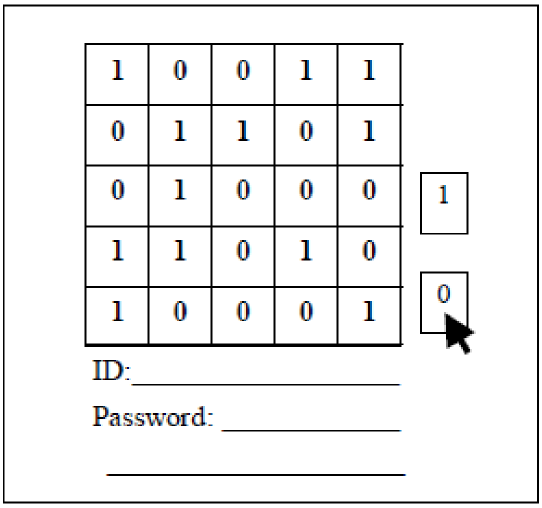
Figure 8. Other forms of login interface

No matter what appears in the grid, the method of the authentication does not change. John's password now is: *V*=[11, 1, 01, 0, 11, 1, 0, 1, 1, 0, 01, 11, 10],or *V*=[aa, a, sa, s, aa, a, s, a, a, s, sa, aa, as].

Although this kind of change would increase the login time for the user, it also increases the security level of the process. Because of that, the length of the text password changes at each login step and the text password space increases.

D.Different Input Style

Because of the high resistant to the shoulder surfing of the keyboard, the input device could be hidden. We can expand the input style of the system by adding the soft keyboard onto the interface. The mechanism can be used in mobile devices or other screen-based input environment. Although the input process can be easily recorded, the scheme has strong resistance to this kind of attack. Fig. 9 shows the example.



1	0	0	1	1
0	1	1	0	1
0	1	0	0	0
1	1	0	1	0
1	0	0	0	1

ID:__________________

Password: _____________

____

Figure 9. Interface without keyboard

IV. Security Analysis

A. Resistant to Shoulder Surfing

Because the login step does not reflect the shape password directly, this hybrid password scheme is highly resistant to shoulder-surfing. In the login step, the only method to obtain the original shape password is that the attackers must record the whole finger process and the certain grid form of that login process. In addition, if a hidden camera has recorded the whole process of the login step, it is not easy to crack the system either.

Take the password used above for example. We suppose that the attacker has obtained the interface and the password number at the login step using a hidden recorder. The interface is shown in Fig. 10 (the same as the basic scheme shown in the previous example) and the password vector is *V*=[1100110110011]. Since the scheme is based on the stroke, the attackers had to guess the accordant stroke with the numbers. It is apparent that one password vector could represent many stroke variants. Figure 9 shows the example of three shapes compared to the former “N”.

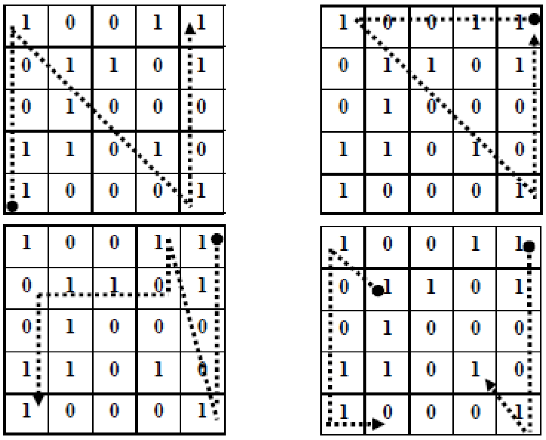


Figure 10. Stroke variants

Because the stroke can be any form of shape, the number of the stroke of one vector is as follows: suppose f_1 means the number of “*1*” and f_2 means the number of “*0*” in the interface, and f_1 = m , f_2 = n , |*V*|= 13 (the size of the shape of V), H_s = m^8n^5 . In this example, f_1 = 12, f_2 = 13, H_s = $12^8 13^5$ = 1.6 × 10^{14}.

B. Resistant to Brute Force

There are two categories of brute force attack: the text-based and the stroke-based.

For the text-based brute force attack, the text space of this password Ht scheme is not large. However, the login passwords required are not the same string in each time when users login. For the string tried by the attackers in parts of the password space may become correct in any time, this method does not have the effect to this scheme.

Another brute force attack is based on the stroke or the shape on the grid. This can be seen more effective than textual-based one. The attacker could try the shape and its stroke on the grid without considering the character appeared in the grid. It is difficult to guess the shape and the stroke of user’s password. Moreover, one shape can have different kinds of stroke sequence, and can have different region in the grid. More importantly, an original password could be composed of several strokes of any shape. All of its mechanism larges the password shape space.

Because the login step does not require user to click on the screen directly, the stroke-based textual scheme is highly resistant to shoulder-surfing. The attackers must record the whole finger process and the certain grid form in order to obtain the password. In addition, if a hidden camera has recorded the two scenes of login, it is not easy to crack the system either.

One type of brutal force attack is trying all of the strokes on the grid, assuming that the attacker does not know any information about the stroke’s *S*, |*S*| and |*V*|. Thus the attackers have to try all of the shapes with any length. The password space Hs is as follows:

$$H_s = g\,|V| \qquad (1)$$

In this example, H_s=25^{13} = 1.5 × 10^{18} .More advanced stroke-based brutal force attack is that the attacker would try the certain stroke, assuming that the attacker has known the shape of the password. For example, the attacker uses the character N to attack the scheme. Without considering the stroke dividing of the same shape, the single shape “N” could have variants, as shown in Fig 11.

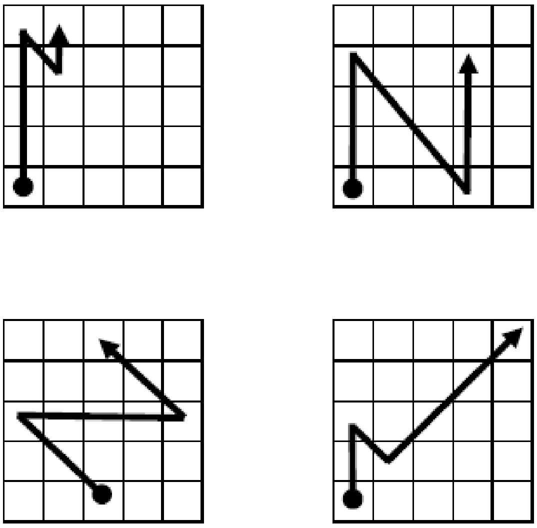


Figure 11. Shape variants of *N*

$H_s = \sum_{m=1}^{|S|} S^* 2^m$, where *S** can only be specified for the certain shape. To this example, we suppose the *S** = 20, |*S*|= 5 , and H_s = 1240 . Although the space is not large enough, it is on the premise that the shape *S* is fixed.

C. Resistant to Random Click Attack

Because the number of kinds of character in the login step is only 2, random clicking with the two keys may become a more effective method to attack the system. Again considering John's password, where $|V|=13$. The possibility of guessing the right string is $1/2^{13} \approx 0.000122$. And if the attacker does not know the length of the password, then the possibility will become less than that. Furthermore, the number of kinds of text appeared in the grid during authentication can be added to 3, thus it is more difficult to get the correct password through random clicking.

D. Multi-step Login

Multi-step login means: users can choose more shapes than one as their original passwords. Different shapes will be inputted according to different interface grid. To avoid increasing the login time, this method is appropriate for the interface with small size grid. In this situation, user's original passwords are a sequence of shaped, can they se and login with them in the same way like the basic scheme.

V. Conclusion and Future work

In this paper, a hybrid password scheme based on shape and text is proposed. The scheme has salient features as a secure system for authentication immune to shoulder-surfing, hidden camera and brute force attacks. It also has variants to strengthen the security level through changing the login interface of the system.

However, the system still has some drawbacks. Firstly, this method is relativity unfamiliar to the general public so that the users may adopt the simple and weak strokes as their passwords. If the shapes chosen by the use have normal meaning, the attacker will have more chance to attack the password. Thus, teaching the user to use this scheme and select the original shape password carefully is very crucial to this new system.

Secondly, the most vulnerable step of this scheme is the password creating step, since the users have to tell the system the original shapes and stokes. If this process is recorded by the attackers, the whole system will be attacked easily. Therefore, proposing a more secure method to replace the set step can enhance the whole system effectively.

Thirdly, the login process is longer than other graphical schemes. And if the input process is not familiar to the user, the text input will be clicked by mistaken if not carefully.

In addition, although some researchers have do some analysis about the memorability of the shape as the password should be investigated deeply and thoroughly [16][17], so that this kind of password scheme can be accepted by the users.

To address these issues, we should design more advanced authentication system to improve this method.

Acknowledgment

This work is carried out under the "Taishan Scholar" project of Shandong China. The research is also supported by the Natural Science Foundation of China (No.60873058, No.60743010), the Natural Science Foundation of Shandong Province (No. Z2007G03), and the Science and Technology Project of Shandong Education Bureau.

The authors wish to thank A, B, C. This work was supported in part by a grant from XYZ.

References

[1] A. Adams and M. A. Sasse, "Users are not the enemy: why users compromise computer security mechanisms and how to take remedial measures," *Communications of the ACM,* vol. 42, pp. 41-46, 1999.

[2] G. E. Blonder, "Graphical passwords," in United States Patent, vol. 5559961, 1996.

[3] S. Wiedenbeck, J. Waters, J. C. Birget, A. Brodskiy, and N. Memon, "Authentication using graphical passwords: Basic results," in *Human-Computer Interaction International (HCII2005)*. Las Vegas, NV, 2005.

[4] W. Jansen, "Authenticating Mobile Device User Through Image Selection," in *Data Security*, 2004.

[5] W. Jansen, "Authenticating Users on Handheld Devices," in *Proceedings of Canadian Information Technology Security Symposium*, 2003.

[6] W. Jansen, S. Gavrila, and V. Korolev, "A Visual Login Technique for Mobile Devices," in *National Institute of Standards and Technology Interagency Report NISTIR 7030,* 2003.

[7] I. Jermyn, A. Mayer, F. Monrose, M. K. Reiter, and A. D. Rubin, "The Design and Analysis of Graphical Passwords," in Proceedings of the 8th USENIX Security Symposium, 1999.

[8] J.Thorpe and P. C. v. Oorschot, "Graphical dictionaries and the memorable space of graphical passwords," in *Proceedings of the 13th USENIX security Symposium,* San Deigo,CA, 2004.

[9] J.Thorpe and P. C. v. Oorschot, "Towards secure design choices for implementing graphical passwords," in *Proceedings of the 20the Annual Computer Security Applications Conference*. Tucson, Arizona,2004.

[10] D. Nali and J. Thorpe, "Analyzing user choice in graphical passwors.," in *Technical Report School of Information Technology and Engineering,* University of Ottawa, Canada, 2004.

[11] H. Zhao and X. Li, "S3PAS: A Scalable Shoulder-Surfing Resistant Textual-Graphical Password Authentication Scheme," in *21st International Conference on Advanced Information Networking and Applications Workshops (AINAW 07),* vol. 2. Canada, 2007, pp. 467-472.

[12] S.Man, D. Hong, and M.Mathews, "A shouldersurfing resistant graphical password scheme," in *Proceedings of International conference on security and management. Las Vergas,* NV, 2003.

[13] D. Hong, S. Man, and B. Hawes, "A password scheme strongly resistant to spyware," in *Proceedings of International conference on security and management*. Las Vergas, NV, 2002.

[14] A. D. Luca, R. Weiss, and H. Hussmann, "PassShape: stroke based shape passwords," in *Proceedings of the 2007 conference of the computer-human interaction special interest group (CHISIG) of Australia on Computer-human*

interaction: design: activities, artifacts and environments. Australia 2007, pp. 239-240.

[15] X. Suo, Y. Zhu, and G. S. Owen, "Graphical passwords: A survey," *21st Annual Computer Security Applications Conference (ASCSAC 2005)*. Tucson, 2005.

[16] R. N. Shepard, "Recognition memory for words, sentences, and pictures," *Journal of Verbal Learning and Verbal Behavior*, vol. 6, pp. 156-163, 1976.

[17] R. Weiss, A. D. Luca, "PassShape – Utilizing Stroke Based Authentication to Increase Password Memorability,” *NordiCHI*, 2008.

Ziran Zheng was born on Decemeber 15, 1981, in Jinan, China. In 2007, he received the Master degree in computer software and theory in Shandong Normal University. Currently he is a PhD student at Shandong Normal University. His major field is computer-aided design, information security.

A New Hybrid Method for Mobile Robot Dynamic Local Path Planning in Unknown Environment

Peng Li, Xinhan Huang, Min Wang
Department of Control Science and Engineering,
Huazhong University of Science and Technology, Wuhan 430074, China
Email: lipeng_hubei@126.com

Abstract—**In this paper, a hybrid approach for efficiently planning smooth local paths for mobile robot in an unknown environment is presented. The single robot is treated as a multi-agent system, and the corresponding architecture with cooperative control is constructed. And then a new method of information fusion namely DSmT (Dezert-Smarandache Theory) which is an extension of the DST (Dempster-Shafer Theory) is introduced to deal with the error laser readings. In order to make A* algorithm suitable for local path planning, safety guard district search method and optimizing approach for searched paths are proposed. Also, the parameters of internal Proportional-Integral-Derivative (PID) controller in the goto agent are adjusted through practical experiments for the use of smoothing the path searched by optimized A* algorithm. Finally, two kinds of experiments are carried out with Pioneer 2-DXe mobile robot: one uses the hybrid method proposed in this paper, the other uses artificial potential field (APF) which is the classical algorithm for local path planning. The experimental results reveal the validity and superiority of the hybrid method for dynamic local path planning. The approach presented in this paper provides an academic support for path planning in dynamic environment with moving objects in the field.**

Index Terms—**path planning, multi-agent, Dezert-Smarandache Theory, A* algorithm, mobile robot.**

I. INTRODUCTION

Mobile robot path planning is one of the most important topics in robotics research. Point-to-point path planning of autonomous mobile robot, defined as finding a collision-free path linking a given start configuration to a goal configuration, has been extensively explored in last two decades. Many different methods achieving varying degrees of success in a variety of conditions/criteria of motion and environments have been developed.

Path planning for mobile robot is composed of two main parts: the global path planning and the local path planning. For the global path planning, the entire environment is known for the robot, so that the robot only needs to compute the path once at the beginning and then to follow the planned path up to the target point. Oppositely for the local path planning, the robot only knows the area which has been detected itself, it only casually decides the direction to move. There are many studies on robot path planning using various approaches, such as the grid-based A* algorithm [1], road maps [2], cell decomposition [3], and artificial potential field (APF) [4]. Some of the previous approaches use global methods to search the possible paths in the workspace [5], normally deal with static environments only, and are computationally expensive when the environment is complex. Some methods suffer from undesired local minima; the robots may be trapped in some cases such as with concave U-shaped barriers. But most studies mentioned above mainly focus on global path planning and there is few valid method for local path planning.

This paper presents an efficient hybrid approach for real time collision-free path planning and grid method is adopted to depict the environment map. In traditional path planning studies, the robot is always treated as a single unit. But here we treat the robot as a multi-agent system, such as path planning agent, behavioral agent and perception agent. In this system, agents can be complex entities at the same time. Each agent achieves its task and collaborates with other agents for the same purpose — to find a rational path. When the robot is commanded to reach an appointed target, in brief, the path planning agent calculates a temporary path with limited knowledge about the surroundings with the A* algorithm and the behavioral agent smoothes the planned path using its goto agent. The error readings are wiped off by filter based on DSmT [6] which is extended from DST [7]. And the path is revised once robot finds an obstacle block the path. In the past, A* algorithm is completely used in global path planning; here it is proved to be a valid algorithm for local path planning, too. And it provides an academic support for path planning in dynamic environment (some objects moves in the environment).

II. MULTI-AGENT ROBOT SYSTEM ARCHITECTURE

It is established that multiple cooperative control in a single robot involves cooperative control and multi-agent systems. Cooperative control has a very general meaning, any complex system developed with multi-agent architectures may be considered as a cooperative approach.

The multi-agent robot system can be divided into five subsystems of agents: perception, behavior, path planning, localization and actuator (Fig.1). The behavioral agent subsystem includes goto agent and avoid agent. In

doi:10.4304/jcp.5.5.773-781

addition to all of the above, there is a client agent acting as user interface. Fig.1 also depicts the flow of information among different agents.

The perception agent obtains information about the environment and about the internal conditions of the robot. Of course, it includes an error reading filter based on DSmT. It collects data from the sensors and after getting rid of error readings, it adapts the data to provide the information requested by the other agents of the system. For example, the milemeter is in charge of obtaining the coordinates (*x*, *y*) of the robot and its orientation, with reference to a fixed axis; the laser sensor collects all the laser readings of the robot.

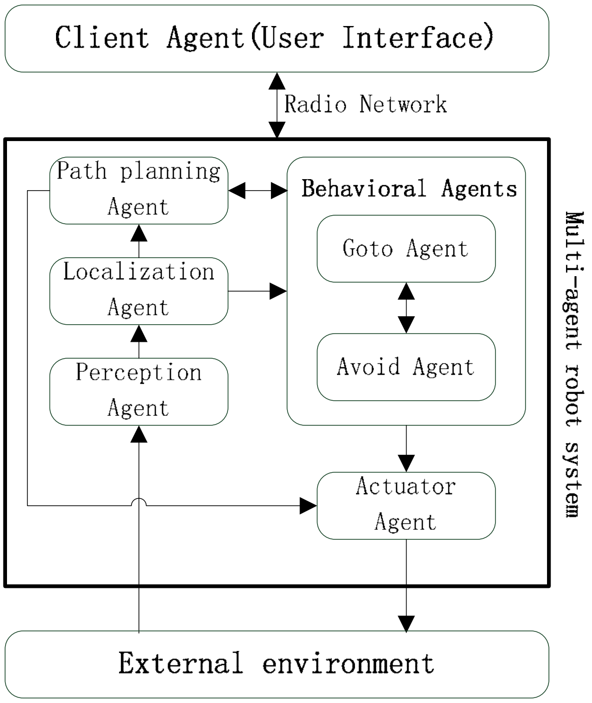


Figure 1. Multi-agent robot system and the relationships among the different agents.

The localization agent locates the robot on the global map, the path planning agent searches for obstacle free paths.

The behavioral agent subsystem carries out specific actions, such as avoiding obstacles, going to a point, etc. The information coming from the localization agent and path planning agent are used to react or respond to the changes produced in the robot itself or in the environment. The following agents have been defined: the goto agent, which is in charge of taking the robot from the initial to the final coordinates without considering obstacles; the avoid agent, which must go around the obstacles when they are found in the path of the robot.

The actuator agent is responsible for directly using the robot's various performance motion components, such as linear and angular velocity controllers, etc.

Once all the agents are running, the user can request a task through the client agent which sends the new robot goal to the path planning agent. Then the path planning agent divides the task into a series of turn point goal and sends the first target position to the goto agent. Based on this information and the actual position (obtained from the localization agent), the goto agent calculates the best linear and angular speeds to reach the target. On the other hand, based on the information provided by the localization agent and laser agent, the avoid agent calculates the linear and angular speeds to dodge the obstacle. At this point both agents (goto and avoid) negotiate in order to decide who uses the motors. But usually the avoid agent does not need to work because the path planning has find a collision free path for robot except accidents, for example, an object abruptly appears in front of the robot and the path planning yet has not calculated the new path for current situation.

The one that wins sends the desired speeds to the actuator agent. And then perception agent obtains the laser and the milemeter readings and sends them to the localization agent, correspondingly. With this new information all the agents update their internal state and new decisions can be taken. Once the target position sent by the path planning agent is reached, the next target position will be sent to the goto agent.

The robot's status can be monitored by client agent. This agent depicts a real-time global map and controls the robot according to the information transmitted from the robot via radio network.

III. ERROR READING FILTER BASE ON DSmT

A. *Simple review of DSmT*

The DSmT of plausible and paradoxical reasoning proposed by the authors in recent years allows to formally combine any types of independent sources of information represented in term of belief functions [8, 9]. And it is mainly focused on the fusion of uncertain, highly conflicting and imprecise sources of evidence. DSmT is able to solve complex static or dynamic fusion problems, especially when conflicts between sources become large and when the refinement of the frame of the problem under consideration, denoted Θ, becomes inaccessible because of the vague, relative and imprecise nature of elements of Θ.

Notion of hyper-power set D^Θ *:* One of the cornerstones of the DSmT is the notion of hyper-power set. Let $\Theta=\{\theta_1,\ldots,\theta_n\}$ be a finite set (called frame) of n elements. The hyper-power set D^Θ is defined as the set of all composite propositions built from elements of Θ with $\cup$ and $\cap$ operators such that:

a) $\phi,\theta_1,\ldots,\theta_n \in D^\Theta$.

b) If $A,B\in D^\Theta$, then $A\cap B\in D^\Theta$ and $A\cup B\in D^\Theta$.

c) No other elements belong to D^Θ, except those obtained by using rules *a)* or *b)*.

Generalized belief functions: We define a map from a general frame Θ $m(.):D^\Theta\to[0,1]$ associated to a given source, say B, of evidence as

$$m(\phi)=0 \quad \text{and} \quad \sum_{A\in D^\Theta} m(A)=1$$

The quantity $m(A)$ is called the *generalized basic belief assignment* (gbba) of A .

The generalized belief and plausibility functions are defined as

$$Bel(A)=\sum_{\substack{B\subseteq A\\ B\in D^{\Theta}}} m(B)$$

$$Pl(A)=\sum_{\substack{B\cap A\neq\phi\\ B\in D^{\Theta}}} m(B)$$

The Classic DSm Rule for Free-DSm Model: For k independent uncertain and paradoxical sources of information providing generalized basic belief assignment $m_i(.)$ over D^{Θ} , the classical DSm conjunctive rule of combination [9] $m_{M^f(\Theta)}(A)$ is given by

$\forall A\neq\phi\in D^{\Theta}$,

$$m_{M^f(\Theta)}(A) □ [m_1\oplus\cdots m_k](A)=\sum_{\substack{X_1,\cdots,X_k\in D^{\Theta}\\ (X_1\cap\cdots\cap X_k)=A}}\prod_{i=1}^{k} m_i(X_i)$$

$m_{M^f(\Theta)}(A)=0$ by definition, unless otherwise specified in special cases when some source assigns a non-zero value to it.

A. Error reading filter base on DSMT

Here the Θ is defined as the status of each grid cell on the map which is constructed by the robot. Suppose there are two elements θ_1 and θ_2 in the frame of discernment Θ. θ_1 means the reading is wrong and θ_2 is defined as right. The hyper-power set is $D^{\Theta}=\{\phi, \theta_1\cap\theta_2, \theta_1, \theta_2, \theta_1\cup\theta_2\}$. Then we define $m(\theta_1)$ as the general basic belief assignment function (gbbaf) for right status; define $m(\theta_2)$ as the gbbaf for the wrong status; $m(\theta_1\cap\theta_2)$ is defined as the gbbaf of conflict mass, it is generated during the fusion; and $m(\theta_1\cup\theta_2)$ is defined as the gbbaf of unknown status (it mainly refers to those areas that still not be scanned at present, so the $m(\theta_1\cup\theta_2)$ of detected areas is zero and does not need join in the fusion).

The laser sensor can detect the 180° area in front of the robot. Each time it can get no more than 180 readings. It is proved that if the surrounding is roomy the amount of readings can reach nearly 180; oppositely if narrow the robot maybe can only get about 150 readings.

There are two evidence sources for the filter. The first source is from the readings themselves. When the robot get a new group of laser readings, each reading is compared with its two neighbor readings (left reading and right reading) except the first and the last reading (because the first one and the last one only have one neighbor reading). For example, if a reading R is being checking, the left reading is defined as R_L and the right reading is R_R, if R□ R_L & R□ R_R or R□ R_R & R□ R_L, these two situations (shown in Fig.2) mean that the reading R probably is a wrong reading and it need to be fused with the second source which is mentioned later. Other situations mean that R is a right reading and does not need further fusion.

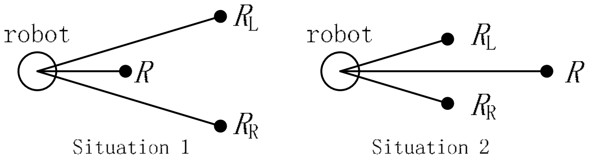


Figure 2. Two situations of error readings.

The belief assignments of the first source $m_1(\cdot)$: $D^{\Theta}\rightarrow[0,1]$ are constructed by authors as follows:

$$R_e=|R_L+R_R-2R|/2; \quad (1)$$

$$R_{max}=\mathrm{Max}\{R, R_L, R_R\}; \quad (2)$$

$$m_1(\theta_1)=\exp[-5\times(R_e/R_{max})^2]; \quad (3)$$

$$m_1(\theta_2)=1-m_1(\theta_1); \quad (4)$$

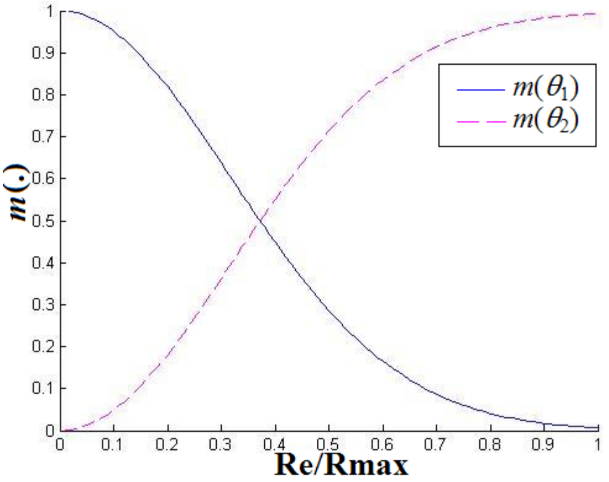


Figure 3. $m(\cdot)$ of Eq.(1~4).

The Fig.3 shows that the smaller the R_e is, the bigger the general basic belief assignment (gbba) of θ_1 becomes, and the gbba of θ_2 is opposite.

The second evidence source is from the map which has been built by the robot. If a laser reading is in one of the two above mentioned error reading situations, the $m_1(\cdot)$ calculated from the first source will be fused with the $m_2(\cdot)$ of second source. The robot checks the area around of the reading point which is marked as the potential error reading. The area is a rectangle with 5×5 grids, and the reading point is the center grid. The $m_2(\cdot)$ of the reading point is calculated according to the amount of occupied grid in this area. Suppose N is the amount of occupied grid in this area, the $m_2(\cdot)$ of the second source is computed as:

$$m_2(\theta_1) = \begin{cases} N/10, & N < 10, \\ 1 \quad , & N \geq 10; \end{cases} \tag{5}$$

$$m_2(\theta_2) = 1 - m_2(\theta_1); \tag{6}$$

The fusion between $m_1(\cdot\ )$ and $m_2(\cdot\ )$ follows the combining rules of Proportional Conflict Redistribution Rule 2 (PCR2) [10] under the framework of DSmT. The PCR2 formula for $k \geqslant 2$ sources is:

$$\forall (X \neq \phi) \in D^{\Theta}, m_{PCR2}(X) =$$

$$[\sum_{\substack{X_1, X_2, \ldots, X_s \in D^{\Theta} \\ X_1 \cap X_2 \cap \ldots \cap X_s = X}} \prod_{i=1}^{s} m_1(X_i)] + C(X) \frac{c_{12\ldots s}(X)}{e_{12\ldots s}} \square k_{12\ldots s} \tag{7}$$

Where

$$k_{12\ldots s} = \sum_{\substack{X_1, \ldots X_s \in D^{\Theta} \\ X_1 \cap \ldots \cap Xs = \phi}} \prod_{i=1}^{s} m_i(X_i) \tag{8}$$

$$C(X) = \begin{cases} 1, & \text{if X involved in the conflict,} \\ 0, & \text{otherwise;} \end{cases}$$

In this formula, $c_{12\ldots s}(X)$ is the non-zero sum of the column of X in the mass matrix, $k_{12\ldots s}$ is the total conflicting mass (here it is $m(\theta_1 \cap \theta_2)$), and $e_{12\ldots s}$ is the sum of all non-zero column sums of all non-empty sets involved in the conflict.

After fusion, the final gbba of θ_1 can be calculated and if $m_1(\theta_1) \geq 0.8$ that means the laser reading which is being checked is a right reading instead of an error reading.

IV. Path planning method

A. Path planning agent

Method of path planning agent: The robot plans the path through path planning agent. After updating the environmental information, the robot firstly makes use of A* algorithm to re-calculate the path if need. The path from current location to the target point is decomposed into a series of turning goal point. And then the real path between every two goal points is smoothed by internal PID controller.

The A* algorithm is usually used for static global path planning. There are few applications for dynamic local path planning in real time because of its large amount of calculations. In this paper, a simple and very effectual method is proposed to make the A* suitable for dynamic local path planning in real time. In most path planning studies, the robot is abstracted as a point without acreage. But actually the robot's radius must be considered in practical applications. The area near the planned path is the safety guard district. That means this area must be empty or the robot cannot pass. The safety guard district is shown in Fig.4.

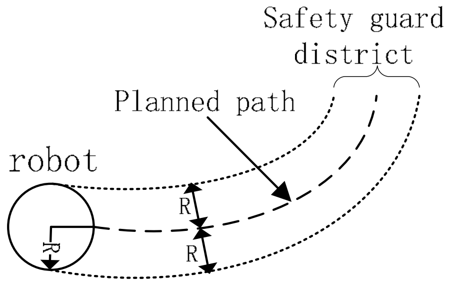


Figure 4. Safety guard district.

Once the robot gets a new group of laser readings, each reading is checked to make sure whether it is in the safety guard district. If a reading is in this area, it means an obstacle blocks the way and the path must be re-calculated. Oppositely, if there is none in the area, well then the path does not need to be re-calculated. This method reduces the computations of path planning. The robot only needs to calculate the path occasionally.

Optimized A algorithm:* The idea of A* algorithm is that each node is associated with a cost function

$$f(n) = g(n) + h(n),$$

where $g(n)$ is the cost from initial node to the current node and $h(n)$ is an estimated cost from the current node to the goal node. A* SEARCH generates and processes the successor nodes in a certain way. Whenever it looks for the next node to process, it employs heuristic function $h(n)$ trying to choose the lowest cost node to process. The following algorithm summarizes A* algorithm:

1. Put the start node *s* on a list called OPEN and compute *f* (*s*).
2. If OPEN is empty, exit with failure; otherwise continue.
3. Remove from OPEN that node whose *f* value is smallest and put it on a list called CLOSED. Call this node *n*. Resolve ties for minimal *f* values arbitrarily, but always in favor of any goal node.
4. If *n* is a goal node, exit with the solution path obtained by tracing back through the pointers; otherwise continue.
5. Expand node *n*, generating all of its successors. If there are no successors, go immediately to 2. For each successor n_i, compute $f(n_i)$.
6. Associate with the successors not already on either OPEN or CLOSED the *f* values just computed. Put these nodes on OPEN and direct pointers from them back to *n*.
7. Associate with those successors that were already on OPEN or CLOSED the smaller of the *f* values just computed and their previous f values. Put on OPEN those successors on CLOSED whose *f* values were thus lowered, and redirect to n the pointers from all nodes whose *f* values were lowered.

Go to 2.

The A* algorithm relies heavily on heuristic function. An appropriate heuristic function determines whether the algorithm can execute efficiently and accurately. In order to find an optimal solution, the heuristic must be admissible. To be admissible, the heuristic function must never over-estimate the cost from one node to the goal node.

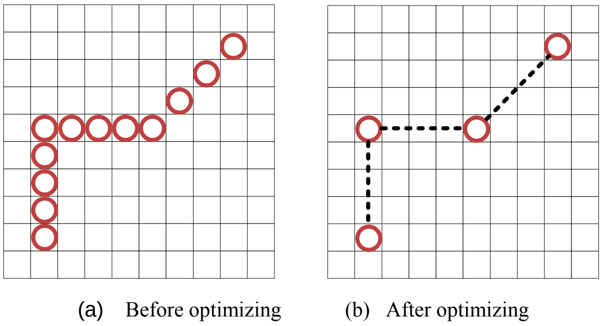


Figure 5. Sketch map of optimizing point-path.

The path searched by A* algorithm is a group of continuous goal points. If the grid is too small, it will spend a large of memory to store the point-path; and if the goal points are placed too closely, the robot yet cannot follow the path well because of the limit of its turning radius. So here the point-path is optimized. The goal points on the same line are deleted and then the robot only needs to store the turning goal points (include start point and end point). This method markedly reduces the memory for storing point-path. The sketch map of optimized point-path is shown in Fig.5.

B. Goto agent

It is known that the turning angle of the path calculated by A* algorithm in the grid map is 45° or 90°. So the path is not smooth and sometimes the robot cannot follow the trajectory because of these stark turnings. The goto agent can solve this problem. The robot's walking between two neighbor goal points is under the charge of goto agent.

Architecture of goto agent: The input of goto agent is the target goal point (x_i,y_i,θ_i) and the output is a group of control parameters (v,ω). v is the robot's velocity and ω is angular velocity for turning. The architecture of goto agent is shown in Fig.6. The variable d in Fig.6 is the distance from current robot's position to the target goal point.

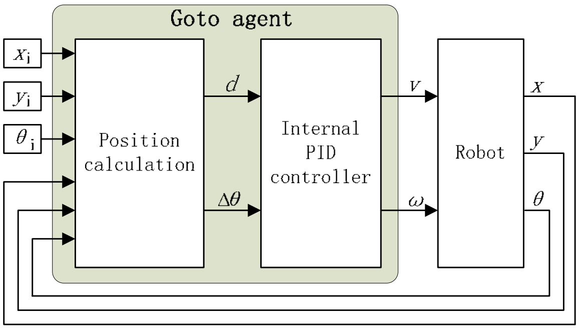


Figure 6. Architecture of goto agent.

Internal PID controller: The robot uses a common Proportional-Integral-Derivative (PID) control system to adjust the PWM pulse width at the motor drivers and subsequent power to the motors.

The P term value K_p increases the overall gain of the system by amplifying the position error. Large gains will have a tendency to overshoot the velocity goal; small gains will limit the overshoot but cause the system to become sluggish. It is found that a fully loaded robot works best with a K_p setting of around 15 to 20, whereas a lightly loaded robot may work best with K_p in the range of 20 to 30.

The D term K_d provides a PID gain factor that is proportional to the output velocity. It has the greatest effect on system damping and minimizing oscillations within the drive system. The term usually is the first to be adjusted if robot encounters unsatisfactory drive response. Typically, it is found that K_d works best in the range of 600 to 800 for lightly to heavily loaded robots, respectively.

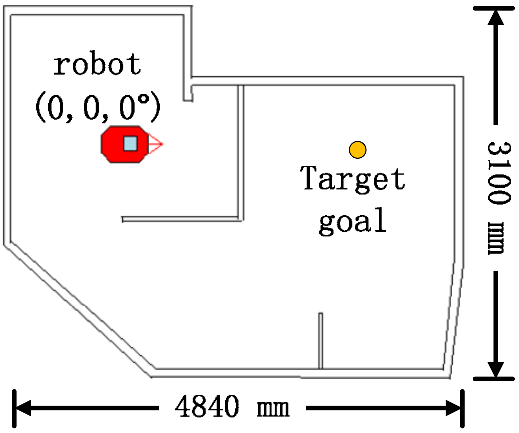
The I Term K_i moderates any steady state errors thereby limiting velocity fluctuations during the course of a move. Too large of a K_i factor will cause an excessive windup of the motor when the load changes, such as when climbing over a bump or accelerating to a new speed. Consequently, this study typically uses a minimum value for K_i in the range of 0 to 10 for lightly to heavily loaded robots respectively.

V. Experiments

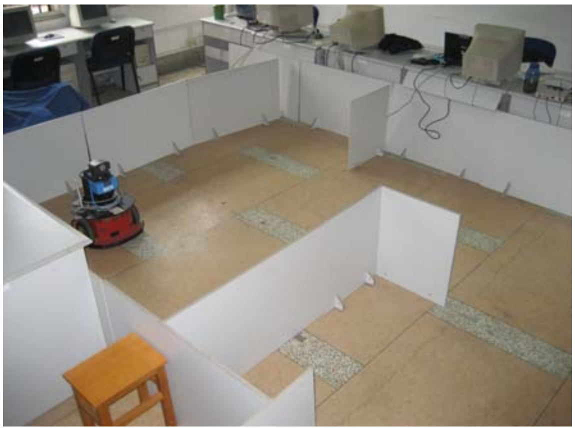
A user interface as a software platform for experiment is developed by authors with Visual Studio 2008. And Pioneer 2-DXe mobile robot which is shown in Fig.7(a) is used in experiments. Two kinds of experiments are carried out: one uses the hybrid method proposed in this paper, the other uses artificial potential field (APF) which is the classical algorithm for local path planning.

An experiment field (size: 4840×3100 mm) is created as Fig.7(b) and the real world of experiment is as Fig.7(c). The point of robot I is treated as the coordinate origin of the global map. So robot is set to the pose of (0,0,0°). The third parameter is the deflection angle of robot. And the target goal position is placed near to the right top corner.

(a) Pioneer 2-DXe mobile robot.



(b) Initial positions of robot.

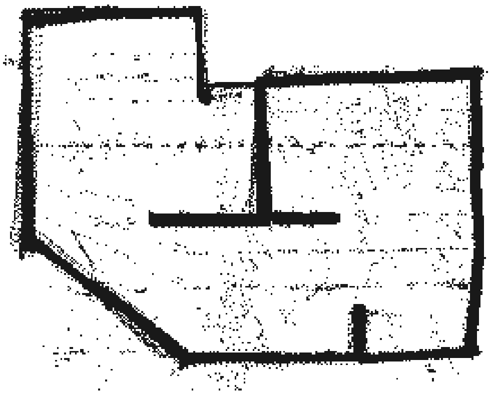


(c) Real world of experiment.

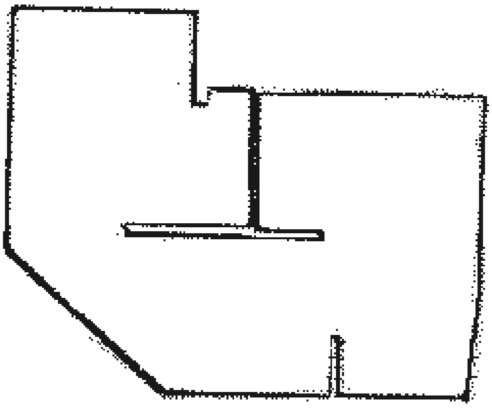
Figure 7. The mobile robot and the experiment field

A. Hybrid method experiment

The effect of error reading filter: The effect of error reading filter is verified before the path planning experiment. The striking dissimilarity between mapping without filter and mapping with filter is revealed clearly in Fig.8. If there is no filter, the robot cannot find the way to the target goal because so many barrier-points block the way that there is not enough space to pass.



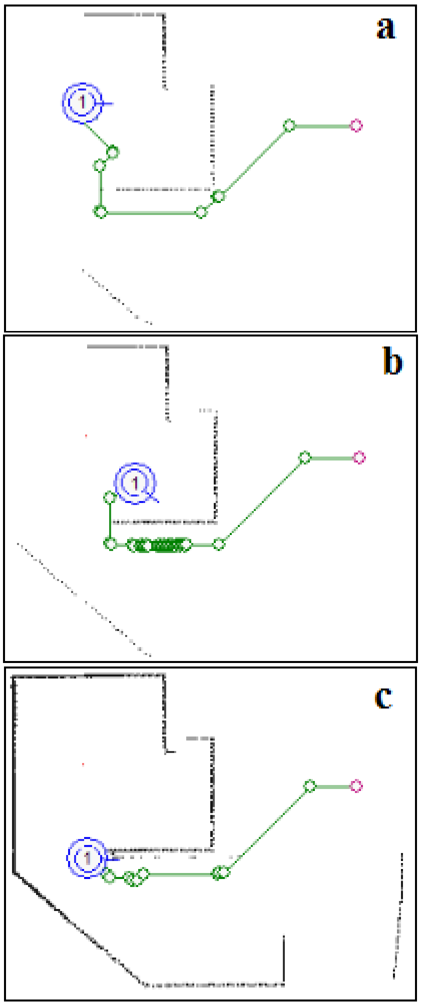
(a) Mapping without filter.



(b) Mapping with filter.

Figure 8. Effect of error reading filter.

Path planning experiment: In order to distinguish the planned path and the final real path, the planned path is displayed during walking and the real path is displayed when the robot arrives at the target goal point. The experimental result is shown in Fig.9. The circles in Fig.9 are the turning goal points calculated by optimized A* algorithm.



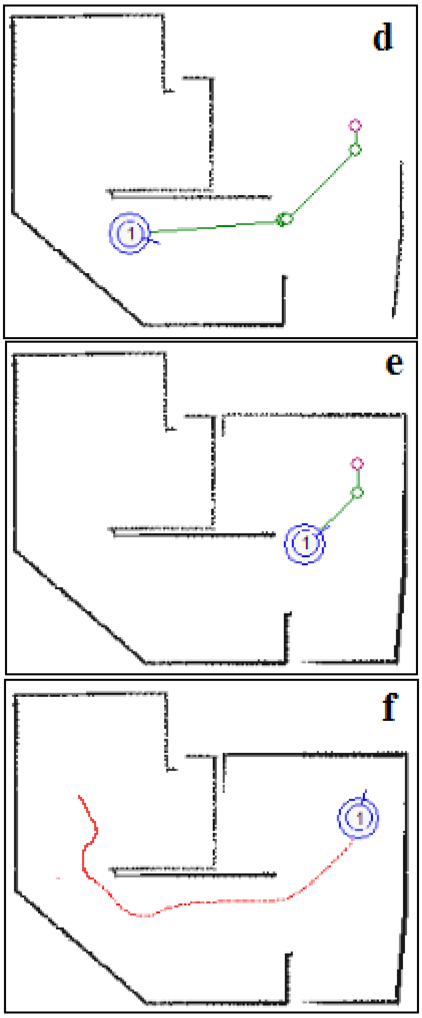


Figure 9. Hybrid method experimental result. (a)~(e): path planning during walking; (f): the real path

B. APF experiment

The error reading filter is still used in this experiment for its important effect. The start position and the target goal position is the same as in the hybrid method experiment. Of course in this experiment the robot is not a multi-agent system any more. For this experiment does not need the behavioral agent subsystem. And then the information flow becomes unilateral. There are no reciprocities between the parts in the robot.

The APF experimental result is shown in Fig.10. The trajectory in the center is robot's real path. The fig.10(h) shows the final path of the robot.

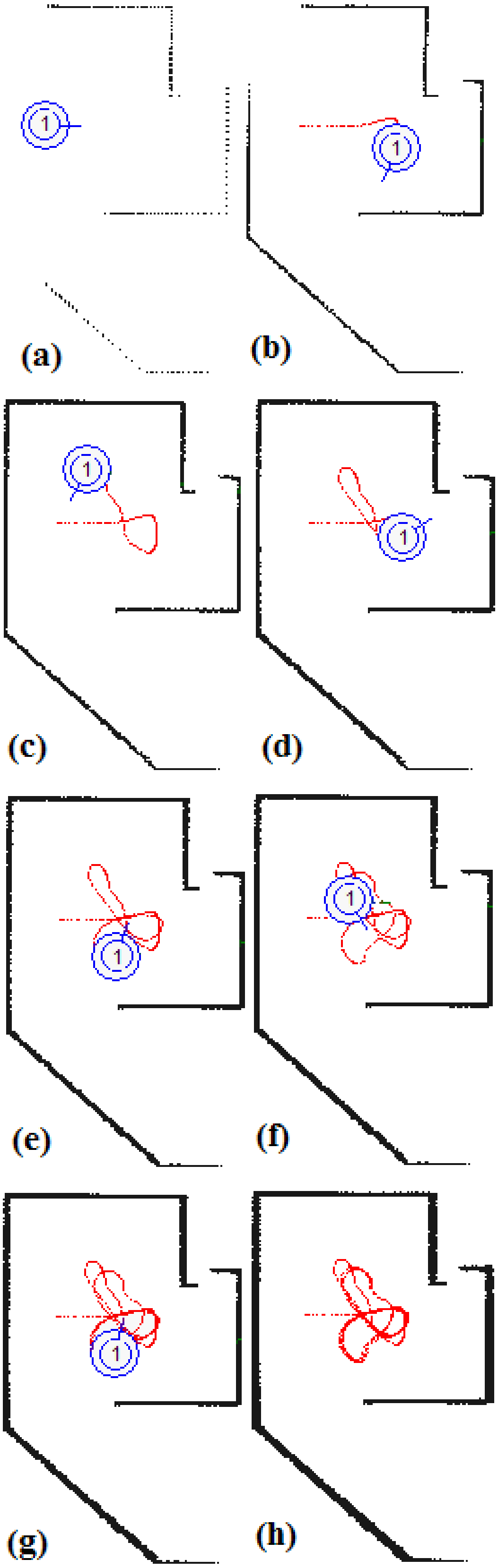


Figure 10. APF experimental result.

C. Analysis

Through these two experiments, it is obviously that the hybrid method proposed in this paper is a very effectual method for dynamic local path planning. The experimental results of these two experiments clearly show:

1) In hybrid method experiment, the robot only needs to store several turning goal points. The planned path is changed once the robot finds new barriers block the way; this is expressly shown in Fig.9(c~d).

2) The goto agent performs so perfectly that the real path of hybrid method experiment is very smooth instead of stark turnings which is the fault of A* algorithm.

3) Actually, the start position of the robot is a concave U-shaped trap. In hybrid method experiment, the robot can easily walk out this area and the path is rational. But in APF experiment, the robot is trapped in local minima; it repeatedly follows the same trajectory and cannot get out the repetition.

4) The effect of error reading filter based on DSmT is valid. The maps built with the laser readings are clean and accurate.

5) In hybrid method experiment, the Safety guard district search method reduces the computation, so that the whole system works well without lags even crash. The experiment is carried out well.

VI. Conclusions

This paper has proposed a hybrid approach for planning smooth paths satisfying dynamic local constraints for robot in an unknown environment. The single robot is divided into several synergic agents; among them the most important agents for path planning are path planning agent and behavioral agents since their cooperation influences directly the final result. This paper also presents a valid error reading filter based on DSmT. And then in order to make A* algorithm suitable for dynamic local path planning, safety guard district search method and optimizing approach for searched path are proposed. The parameters of internal PID controller in the goto agent are adjusted through practical experiments. The results of experiments carried out with Pioneer 2-DXe mobile robot prove the validity and superiority of the hybrid method for dynamic local path planning. The application of the approach presented in this paper to path planning in completely dynamic unknown environment with moving objects around the robot will be investigated in the future.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under grant No. 60675028, which is greatly acknowledged.

References

[1] Chia Hsun Chiang, Po Jui Chiang, Fei, J.C.-C., et al. "A comparative study of implementing Fast Marching Method and A* SEARCH for mobile robot path planning in grid environment: Effect of map resolution", *Proc. IEEE Workshop Adv. Rob. Soc. Impacts, ARSO.* Hsinchu, Taiwan, Dec 9-11, 2007. pp. 1 – 6.

[2] Jung Sungwon, Pramanik Sakti. "An efficient path computation model for hierarchically structured topographical road maps", *IEEE Trans Knowl Data Eng*. 2002, 14 (5): 1029-1046.

[3] Rosell Jan, Iniguez Pedro. "Path planning using harmonic functions and probabilistic cell decomposition", *Proc IEEE Int Conf Rob Autom*. Barcelona, Spain, Apr 18-22, 2005. pp. 1803-1808.

[4] Zhu Qidan, Yan Yongjie, Xing Zhuoyi. "Robot path planning based on artificial potential field approach with simulated annealing", *Proc. ISDA Sixth Int. Conf. Intelligent Syst. Design Applic*., Jinan, China, Oct 16-18, 2006. pp. 622-627.

[5] Z. X. Li and T. D. Bui, "Robot path planning using fluid model," *J. Intell. Robot. Syst.*, 21(1), pp. 29–50, Jan. 1998.

[6] Dezert J. Foundations for a new theory of plausible and paradoxical reasoning. *Information and Security,* 2002, 9: 13~57.

[7] Shafer G. "*A mathematical theory of evidence*", Princeton. N.J: Princeton University press. 1976.

[8] Dezert J, Smarandache F. "On the generation of hyper-powersets for the DSmT", *Proceedings of the Sixth International Conference on Information Fusion, 8-11 July 2003, Cairns, Queensland, Australia*, pp. 1118-1125.

[9] Dezert J, Smarandache F. "*Advances and Applications of DSmT for Information Fusion Vol.1"*. Rehoboth: American Research Press, 2004.

[10] Dezert J, Smarandache F. "*Advances and Applications of DSmT for Information Fusion. Vol.2"*. Rehoboth: American Research Press, 2006.

Peng Li, born in 1981. Since 2006, he has been a Ph.D. candidate in Department of Control Science and Engineering of Huazhong University of Science and Technology (HUST), Wuhan, China. His main interests include information fusion, robot's map building and localization, multi-robot system.

Xinhan Huang, born in 1946. He is currently a Professor, Ph.D. candidate supervisor of Department of

Control Science and Engineering and Head of the Institute of Intelligent Control of HUST. His research interests include robots and their intelligent control, multi-sensor data fusion, pattern recognizing.

Min Wang, born in 1954. She is currently a Professor of the Department of Control Science and Engineering of HUST. Her research interests include robotics, intelligent control, image processing, pattern recognizing.

A Reliable Fuzzy Theory based Reputation System in Grid

Liao Hongmei
School of Computer Science and Technology China University of Mining and Technology, Xuzhou, China
Email: lhm@cumt.edu.cn
Wang Qianping and Li Guoxin
School of Information and Electric Engineering, China University of Mining and Technology, Xuzhou, China
Email:qpwang@cumt.edu.cn

Abstract—**Trust is a fundamental concern in Grid environment. Behavior trust that varies with time is indispensable in the trust system. Because of the fuzzy nature of behavior trust, it is more appropriate to adopt fuzzy logic to express and compute trusts and reputations than adopt probabilities approach. A new reliable fuzzy theory based reputation system in Grid environment is proposed. By variable weighted fuzzy comprehensive evaluation, Direct Trust can be gotten and the affection of the ill factor have low value can be highlighted by give it a larger weight and then Grid entities can avoid choosing service providers with some defective. By derivation and combination of trust, Reputation can be obtained. Expert's experience is used to set and simplify fuzzy rules. A distributed hash table based Dual Two-Layer Chord Protocol is proposed to store and retrieve reputations. Simulation results show that entities in Grid can use resources or deploy services more securely in the support of the reliable fuzzy theory based reputation system.**

Index Terms—**trust model, Grid, fuzzy comprehensive evaluation, reputation system, distributed hash table**

I. INTRODUCTION

Security has been the focus of grid systems recently. Trust model plays an important role in security field. It is often classified into two categories: identity trust and behavior trust. Identity trust is static. Once the identity is authenticated, the behaviors of the entities will not be monitored any more even though they might do something harmful, whereas behavior trust is dynamic trustworthiness. Behavior trust is based on transactions between entities in the past time. If the entity do something wrong or harmful, its behavior trust value will then be dropped down, other entities may decide to give up choosing it as a service provider based on its behavior trust value.

Many trust models based on entity behaviors have been proposed. In the trust model proposed by Alfarez Abdul-Rahman and Stephen Hailers, trust is divided into Direct Trust and Recommend Trust [1]. Audun Jφsang, Andrew Whitby [2] [3] proposed a beta reputation system, which is based on beta probability density functions with which feedback and derive reputation ratings are combined. Lik Mui, Mojdeh Mohtashemi [4] proposed a rating system based on a Bayesian probabilistic framework which also used the beta distribution. But trust is a subjective and inaccurate value which is decided by the Grid entity, it is difficult to describe with accurate probability distribution. It is also difficult to ensure the independency of events in beta reputation system and Bayesian probabilistic framework.

Few behavior trust models based on fuzzy logic in Grid are proposed [5] [6] [7]. But the fixed weighted fuzzy comprehensive evaluation in [5] is not suited; Shanshan Song and Kai Hwang suggest enhancing the trust index of a resource site by upgrading its intrusion defense capabilities and checking the success rate of jobs running on the platforms, but the computing of directed trust is not mentioned in [7].

In this paper, a new reliable fuzzy theory based reputation system in Grid is proposed. Direct Trust is gotten by variable weighted fuzzy comprehensive evaluation and expert experience is used to set and simplify fuzzy rules while used in computing of reputation. By variable weight based fuzzy comprehensive evaluation, the affection of the ill factor have low value can be highlighted by give it a larger weight and Grid entities can avoid choosing service providers with some defective.

A reputation system gathers, distributes, and aggregates feedback about participant's behaviors and it can help people make decisions about who to trust. Most literatures stored reputation information in the node itself or in some central nodes. However, storing in the node itself is not secure enough and Storing in central nodes is easy to cause bottleneck. To address this problem, we proposed a Distributed Hash Table(DHT) based Dual Two-Layer Chord Protocol (DTLCP) as reputation storage and retrieval infrastructures with which the reputation system can provide efficient storage and queries that operate in O(logN) +O(M)overlay hop(N is the number of domains in Grid and N the number of nodes in a domain, M<<N) and can answer queries even if the system is continuously changing. Initialization of reputation and updating of reputation are all discussed in this paper.

Corresponding author: Liao Hongmei .

doi:10.4304/jcp.5.5.782-790

II. TRUST AND REPUTATION

In this paper, we will adopt the definition of trust in papers [8] [9].

Trust is usually divided into direct and indirect trust. If an entity *P* did not transact with entity *Q* in the past, it has no direct trust value about *Q*. An entity without direct trust or having no enough confidence on its direct trust on another entity should use indirect trust. Recommend trust and reputation are both indirect trust.

A Definition of direct trust and reputation

The definition of direct trust and reputation are as follows:

Trust is the firm belief in the competence of an entity to act as expected such that this firm belief is not a fixed value associated with the entity but rather it is subject to the entity's behavior and applies only within a specific context at a given time.

The direct trust level on an entity is built on past experiences and is given for a specific context and a given time frame. For example, entity *P* might trust entity *Q* to use its storage resources but not to execute programs using these resources. The trust level is specified for a given time frame because the trust level today between two entities is not necessarily the same trust level a year ago.

The definition of reputation: The reputation of an entity is an expectation of its behavior based on other entities' observations or the collective information about the entity's past behavior within a specific context at a given time.

Seeking the reputation of a specific entity relies on information from a set of other entities referred to as recommenders set. A recommender is an entity that gives recommendation using its direct trust table that includes trust values for entities with which the recommender had prior direct transactions.

Usually a node *P* need to integrate direct trust stored itself and indirect trust such as reputation to determine whether or not trust entity *Q* and transact with it. The final trust value is computed as follows:

$\alpha *$directly trust $\oplus$ $\beta *$ indirect trust(reputation)

α and β respectively expressed the weights of these two trust relationship in trust system. When the interactions between entities *P* and *Q* frequently, directly trust should get a larger weight. Set $\alpha>\beta$, that is, trustworthiness of node *Q* is based more on direct relationship with ***Q*,** rather than the reputation of ***Q*.** While when the entities *P* and *Q* have little or no contact with each other a long time, set $\alpha<\beta$.

B. Trust Relationships Within domain and Trust Relationships between domains

A Grid computing system is a geographically distributed environment with autonomous domains that share resources amongst themselves. One primary goal of such a Grid environment is to encourage domain-to-domain interactions and increase the confidence of domains to use or share resources: (a) without losing control over their own resources; and (b) ensuring confidentiality for others. To achieve this, the trust relationship can be divided into in-domains trust and cross-domain trust. Cross-domain trustworthiness makes such geographically distributed systems become more attractive and reliable for day-to-day use. Monitoring and managing the behavior of the entity and building a trust level based on that behavior is needed.

In-Domain trust relationship is the trust relationship between nodes in the same Grid domain, usually decided by both domain strategy and behavior trust. It is more stable and easy to confirmed; strategies such as identity trust can be used simultaneously. In-domain trust relationship is often centralized management, our paper focuses on the more complex and common cross-domain trust relationships. Notation trust and reputation represent the cross-domain trust and cross-domain reputation seperately in the rest of this paper.

Cross-domain trust is defined as trust relationship of Grid nodes in different Grid domains in this paper. Some authors get the final trust value of node *P* on *Q* in different domains by combining the weight of node *Q* in domain D_Q that *Q* is exist and domain trust of D_Q. Domain trust is determined by transaction history of all nodes in this domain.

However, the weight of node within domain can not reflect its performance outside. And if there is a malicious node in domain, according to the penalty strategy, the domain trust will be greatly reduced, so that nodes with high credibility in this domain will lose many trading opportunities.

In this paper, if entity *P* and entity *Q* are cross-domain nodes, the final trust value of *P* on *Q* will depend on their mutual direct trust and reputation of node *Q* also.

III. RELIABLE FUZZY COMPREHENSIVE EVALUATION OF TRUST

Because trust is determined by multiple factors such as an entity's capabilities, honesty and reliability, and so on, we will first give a reliable comprehensive evaluation of direct trust.

With fuzzy theory, trust can partially belong to a set and this is represented by the set membership.

Let $X=\{x_0,x_1,\cdots\cdots,x_n\}$ be the problem domain that trust manage will be researched in. x_i (*i=1, 2,..., n*) denote the entity in the Grid. The definition of fuzzy set is:

A fuzzy set is any set that allows its members to have different grades of membership (membership function) in the interval [0, 1].

Definition: Let *X* be the domain, and let *x* be the element of the set *X*, $\forall x \in X$ there is the mapping:

$$X \to [0,1],\ x \mapsto \mu_A(x) \in [0,1]$$

The fuzzy set *A* in *X* is expressed as a set of ordered pair $A=\{(x \mid \mu_A(x))\},\ \forall x \in X$.

$\mu_A(x)$ is the membership function of fuzzy set *A*, which describes the membership of the element *x* of the base set *X* in the fuzzy set *A*. The grade of membership $\mu_A(x_0)$ of a membership function $\mu_A(x)$ describes for the special element $x=x_0$, to which grade it belongs to the fuzzy set A. This value is in the unit interval [0, 1].

In Grid, the grade of trust can be described by membership degree of different fuzzy sets in X which denote different trust levels. M different fuzzy sets T_i $(i=1, 2,..., M)$, which are in the set of all fuzzy sets in X can be used to denote M different trust levels. For example, when $M=6$, six fuzzy sets T_i $(i=1, 2,...,6)$ can be used to denote six different trust levels in Grid. The trust level of T_i $(i=1,2,...,6)$ is defined as follows:

T_1 denotes the “entire trust” fuzzy set;
T_2 denotes the “very trust” fuzzy set;
T_3 denotes the “trust” fuzzy set;
T_4denotes the “distrust” fuzzy set;
T_5 denotes the “very distrust” fuzzy set;
T_6 denotes the “entire distrust” fuzzy set.

The membership function of x to fuzzy set T_j can be denoted as $T_j(x)$. To a concrete x_i, the membership degree is T_j (x_i), which can be marked as T_{ij}. Trust vector of x_i is :
$V=\{v_0,v_1,...,v_6\}$,v_j $(j=1,2,...,6)$, v_j denotes the membership degree of x_i to T_j. But x_i can simultaneously belong to another fuzzy set T_k, such that T_k (x_i) characterizes the grade of membership of x_i to T_k.

A. Ordered Weighted Averaging Operations

The discussion above does not take the context of trust and different factors into account in trust computing. In real Grid environment, Trust and Reputation both depend on some context. For example, entity P trusts entity Q as multimedia provider, but it does not trust Q as a storage provider. So in the context of requesting a multimedia service, Q is trustworthy. But in the context of providing storage service, Q is untrustworthy.

Even in the same context, the trust is not determined by only one factor. It is often evaluated from several aspects (factors). For each aspect, it develops a kind of trust. The overall trust depends on the combination of the trusts in each aspect. Every factor play role in deciding whether the entity is trustworthy to interact with in this context.

For example, as for the total trust evaluation in some context, there are four important factors (attributes). Suppose the following is the evaluation factors (attributes) aggregation:

$$E=\{ E_1,E_2,E_3,E_4\}$$

E_1,E_2,E_3,E_4 represent service time, contract abidance, stipulate violating and job success rate separately. There may be several subgroup factors in each factor. In this paper, for simplification, we will not discuss the subgroup factors.

Suppose the Trust fuzzy evaluation aggregation is:

$$T=(T_1,T_2,T_3,T_4,T_5,T_6)$$

The relationship R denotes the mapping of E to T. That is, the element r_{ij} in R is the membership degree of e_i to T_j.

Matrix $R=(r_{ij})_{4\times6}$ is the fuzzy comprehensive judgment Matrix of entity trust in some context.

The final trust value will be combined closely with the value assignment of each evaluation factor in E. The nature of weight is shown in the quantity of different factors on objects at different levels, i.e. the different influence from all aspects on the Trust. Let $w=(w_1,w_2,...,w_n)$ be a "weighting vector" such that w_i is in [0, 1] for all i and $\sum_{1}^{n} w_i = 1$.

Suppose w_i is the weight of the factor e_i. So, the trust vector $V=\{v_1,v_2,v_3,v_4,v_5,v_6\}$can be gotten by the following fuzzy mapping:

$$\{v_1,v_2,v_3,v_4,v_5,v_6\}=(w_1,w_2, w_3,w_4) \circ (r_{ij})_{4\times6} \text{ And}$$
$$v_i=w_1*r_{1j}+ w_2*r_{2j}+ w_3*r_{3j}+ w_4*r_{4j} \quad (j=1,2,\ldots,6)$$

This aggregation operation is often called "ordered weighted averaging operations".

B. Varialble Weight based Fuzzy Comprehensive Evaluation of Trust

In ordered weighted averaging operations, weighting vector $w=(w_1,w_2,...,w_n)$ is fixed. That is, once the weight of each factor is set, it will not be changed whatever the actual situation is.

But if the value of some attribute of an entity such as contract abidance is too small, even if the other attributes are all good, the entity is not trusty. In ordered weighted averaging operations, the high value of the other attributes will counteract the low value of the bad attribute contract abidance. If we increase the weight of contract abidance, when the entity abide contract well, other attributes can not be manifested.

By increasing the weight of the attribute when its value is low, we can give prominence to deficiency. That is variable weight based fuzzy comprehensive evaluation of trust which can be described as follows:

1) Suppose the value of evaluation factors (attributes) $E=\{E_1,E_2,E_3,E_4\}$ is denoted as $u_1,u_2,...,u_n$. $u_i\in[0, u_m]$, $i=1,2,...,n$. u_i is the trust value of factor Ei .When attribute E_i is in the best situation , $u_i = u_m$. When attribute E_i is in the worst situation , $u_i = 0$.

2) The weight of factor E_i is denoted as w_i, it is function of $(u_1,u_2,u_3,...,u_n)$. That is, $w_i=w_i(u_1,u_2,...,u_n)$, $(i=1,2,...,n)$. The weight of factor Ei , depends on the separate values of all factors. $w_i\in[0,1]$ and $\sum_{1}^{n} w_i = 1$.

Especially, $w_{mi}=w_i(u_m, u_m, ..., u_m)$, $i=1,2,...,n$ and $w_{mi} \in (0,1), \sum_{i=1}^{m} w_{mi} = 1$. w_{mi} is called base weight , and it can be gotten by hierarchical analytic approach.

$w_{0i}=w_i(u_m ,..., u_m , 0 , u_m ,..., u_m)$, $i=1, 2, ... , n$, $w_{0i}\in(0,1)$. w_{0i} denotes the weight of factor E_i when E_i has its min value and the other factors have their max values. w_{0i} can be set by the specialist as the max value of weight of factor E_i. So the influence of factor E_i can be amplified.

3) In order to find appropriate variable weight $w(u_1,...,u_i,...,u_n)$, $i=1,2,...,n$, which is nonincreasing function of u_i and $\sum_{i=1}^{n} w_i u_i$ which is nondecreasing function of u_i, we introduce function $\lambda_i(u)(i=1,2,...,n)$, which matches the following criteria：

- $\lambda_i(u)$is defined in $[0,u_m]$，it is not negative and bounded；

- u_i is a nonincreasing differentiable function in $[0,u_m]$; $\lambda_i'(u) \le 0, (u \in [0, u_m])$;
- let $\lambda_i(0)= \lambda_{0i}$, $\lambda_i(u_m)= \lambda_{mi}$. λ_{0i} denotes the maximum of $\lambda_i(u)$ and λ_{mi} denotes the minimum value of $\lambda_i(u)$.

To a given set of single-factor assessment of $(u_1,\dots,u_i,\dots,u_n)$, a function $\lambda_i(u_i)(i=1,2,\dots,n)$ can be obtained by the means in literature[10], let

$$w_i(u_1,u_2\dots,u_n)=\frac{\lambda_i(u_i)}{\sum_{j=1}^{n}\lambda_j(u_j)} \quad (i=1,2,..,n) \tag{1}$$

It have been proved that $w_i(u_1,u_2,\dots,u_n),i=1,2,\dots,n,$ in (1) can be the variable weight [10].

C. Application of Variable Weight based Fuzzy Comprehensive Evaluation in Trust Model

Suppose in the trust model node *P* needs to give a comprehensive evaluation of another node *Q* which have had trade with it in the context of storage services.

The trust value in storage services of node *Q* is determined by four factors (attributes) ,that is ,the set of evaluation factors is {compliance with the contract, the completion time of service, quality of service, whether or not a malicious act}, their basis weights have been known, which are (0.3,0.2,0.3,0.2), the max weights of these four factors $(w_{01},w_{02},w_{03},w_{04})$ are determined by some experts, which are (0.7,0.5,0.6,0.7).Let

$$\lambda_{mi}=w_{mi}(i=1,2,\dots,n) \tag{2}$$

$(\lambda_{m1}, \lambda_{m2}, \lambda_{m3}, \lambda_{m4})$= (0.3, 0.2, 0.3, 0.2). By definition of w_{0i} and w_{mi}, as well as (1)and (2) ,we can get formula (3):

$$w_{0i}=\frac{\lambda_{0i}}{\lambda_{0i}+\sum_{j\ne i}w_{mj}}, i=1,2,\dots n \tag{3}$$

$$\lambda_{0i}=\frac{w_{0i}\sum_{j\ne i}w_{mj}}{1-w_{0i}}, i=1,2,\dots, n \tag{4}$$

By calculation, we can obtain that $(\lambda_{01}, \lambda_{02}, \lambda_{03}, \lambda_{04})$ = (1.63, 0.8, 1.05, 1.87). .

For given $u_1,u_2,\dots,u_n$, Fixed i ,let u_i changed to be u, and $u>=u_i$ .Denote

$$\lambda_0=\sum_{j\ne i}\lambda_j(u_j), v_0=\frac{1}{\lambda_0}\sum_{j\ne i}\lambda_j(u_j)u_j \tag{5}$$

$$w_i(u)=w_i(u_1,\dots,u_{i-1},u,u_{i+1},\dots,u_n) \tag{6}$$

It can be testified that:

1) $w(u_1,\dots,u_i,\dots,u_n)$ is a nonincreasing function of u_i unconditionally;

2) Suppose $w_i(u_1,u_2,\dots,u_n)$ have been obtained by (1) ,the necessary and sufficient conditions of that $\sum_{j=1}^{n}w_ju_j$ is a nondecreasing function of u_i are:

$$\lambda'_i(u)\ge -\frac{1}{\lambda_0(u-v_0)}\lambda_i(u)(\lambda_0+\lambda_i(u)) \tag{7}$$

There are usually three solution to compute the λ_i which meets (7) [10], In this paper ,we choose the following formula [10].

To a fixed i, there is:

$$\lambda_{*i}=\sum_{j\ne i}w_{mj}\le\lambda_0\le\sum_{j\ne i}\lambda_{0j}=\lambda^*_{\bullet i} \tag{8}$$

By expression (8), the values of $\lambda_{*i}, \lambda^*_{\bullet i}$ ($i=1,2,\dots,n$)can be calculated for each identified i.

By expression (7), fix i, let

$$\begin{cases}\dfrac{d\lambda_i(u)}{du}=-\dfrac{u_m^{k_i-1}}{\lambda^*_{\bullet i}u^{k_i}}\lambda_i(u)(\lambda^*_{\bullet i}+\lambda_i(u))\\ \lambda_i(0)=\lambda_{0i},\lambda_i(u_m)=w_{mi}\end{cases} \tag{9}$$

It can be proved that $\lambda_i(u\ )$derived from (9) and $w_i(u_1,u_2,\dots,u_n)$ determined by (1) need to meet that $\sum_{j=1}^{n}w_ju_j$ is a nondecreasing function of u_i [10]. The solution of (9) can be gotten as follows:

$$\lambda_i(u)=\frac{\lambda^*_{\bullet i}\lambda_{0i}}{\lambda^*\exp(\frac{1}{1-k_i}(\frac{u}{u_m})^{1-k_i})}, i=1,2,\dots, n \tag{10}$$

In (10),

$$\lambda^*=\sum_{i=1}^{n}\lambda_{0i}=5.35$$

, and

$$k_i=1-\frac{1}{\ln\frac{\lambda_{0i}(\lambda^*_{\bullet i}+w_{mi})}{\lambda^*w_{mi}}}.$$

Solution $w_i(u_1,u_2,\dots,u_n)$ $(i=1,2,\dots,n)$ can derived from (1)an (10).

According to the formulas (3) , (8) and (10), λ_{0i}, λ_{*i} , $\lambda^*_{\bullet i}$ and k_i , for specific i ,can be calculated as shown in Tab. Ⅰ.

TABLE I.
VARIABLES IN VARIALBLE WEIGHT BASED FUZZY COMPREHENSIVE EVALUATION OF TRUST

	E_1	E_2	E_3	E_4
w_{mi}	0.3	0.2	0.3	0.2
λ_{0i}	1.63	0.8	1.05	1.87
$\lambda^*_{\bullet i}$	3.72	4.55	4.3	3.48
λ_{*i}	0.7	0.8	0.7	0.8
k_i	0.289	0.211	0.09233	0.4627

Suppose there are two sets of trust evaluation of factors$\{E_1,E_2,E_3,E_4\}$ whose define area is [0, 10] . The values of these two sets of trust evaluation are respectively [6,8,7,9] and [7,9,8,4].

To get the fuzzy comprehensive trust evaluations of these two groups, we should first change the trust evaluation on [0,10] into membership degrees in six fuzzy sets. The collection of trust evaluation reviews is described in section Ⅱ. Relation *R* represents the map from *E* to *T*. Element r_{ij} in *R* represents the membership degree of e_i to fuzzy set T_j.

$R=(r_{ij})_{4*6}$ denotes the fuzzy comprehensive evaluation matrix in some context. If these six fuzzy sets T_1,T_2,T_3,T_4,T_5,T_6 all adopt trapezoidal membership functions, and parameters [a, b, c, d] are set separately as [-1.8, -0.2, 0.2, 1.8], [0.2, 1.8, 2.2, 3.8], [2.2,3.8, 4.2, 5.8], [4.2, 5.8, 6.2, 7.8], [6.2, 7.8, 8.2, 9.8], [8.2, 9.8, 10.2, 11.8], then the first set of trust evaluation [6,8,7,9] can be fuzzyfied and changed into a fuzzy comprehensive evaluation matrix as shown in Fig. 1:

$$\begin{array}{c|cccccc|}
 & T_1 & T_2 & T_3 & T_4 & T_5 & T_6 \\
E_1 & r_{11} & r_{12} & r_{13} & r_{14} & r_{15} & r_{16} \\
E_2 & r_{21} & r_{22} & r_{23} & r_{24} & r_{25} & r_{26} \\
E_3 & r_{31} & r_{32} & r_{33} & r_{34} & r_{35} & r_{36} \\
E_4 & r_{41} & r_{42} & r_{43} & r_{44} & r_{45} & r_{46}
\end{array} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 \end{bmatrix}$$

Figure 1. Fuzzy Comprehensive Evaluation Matrix of Trust with Evaluation Vector is [6, 8, 7, 9].

The second set of trust evaluation [7, 9, 8, 4] is changed into a fuzzy comprehensive evaluation matrix as shown in Fig. 2:

$$\begin{array}{c|cccccc|}
 & T_1 & T_2 & T_3 & T_4 & T_5 & T_6 \\
E_1 & r_{11} & r_{12} & r_{13} & r_{14} & r_{15} & r_{16} \\
E_2 & r_{21} & r_{22} & r_{23} & r_{24} & r_{25} & r_{26} \\
E_3 & r_{31} & r_{32} & r_{33} & r_{34} & r_{35} & r_{36} \\
E_4 & r_{41} & r_{42} & r_{43} & r_{44} & r_{45} & r_{46}
\end{array} = \begin{bmatrix} 0 & 0 & 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Figure 2. Fuzzy Comprehensive Evaluation Matrix of Trust with Evaluation Vector is [7,9,8,4].

Adopt fix weighted fuzzy comprehensive evaluation; let $w=(w_1,w_2,\ldots,w_n)$ represents the fixed weight vector of factor vector $(E1,E2,E3,E4)$.$w_i\in[0, 1]$ stands for the weight of factor e_i, $\sum_{1}^{n} w_i = 1$. By the following fuzzy mapping:

$\{v_1,v_2,v_3,v_4,v_5,v_6\}=(w_1,w_2, w_3,w_4)\ o\ (r_{ij})_{4\times 6}$ And $v_i=w_1*r_{1j}+ w_2*r_{2j}+ w_3*r_{3j}+ w_4*r_{4j}$ (j=1,2,…,6), comprehensive trust vector $V=\{v_1,v_2,v_3,v_4,v_5,v_6\}$ can be gotten.

The comprehensive value of the two sets of trust evaluation can be separately computed:

V_1= [0 0 0 0.45 0.45 0.1];

V_2= [0 0 0.2 0.15 0.55 0.1].

Defuzzification is the reverse process of fuzzification. Trust and Reputation we have gotten above is a "fuzzy" result, that is, the result is described in terms of membership in fuzzy sets. Defuzzification would transform this result into a single number indicating the trust level of an entity. This may be necessary if we wish to output a real number to the user. An average of maxima method or a centroid method can be used to do this work.

It is usually assumed that an entity P will only engage in a transaction with entity Q if the level of trust exceeds some personal threshold (the level of acceptable trustworthiness), which depends on the transaction context.

Defuzzification is the reverse process of fuzzification. Trust and Reputation we have gotten above is a "fuzzy" result, that is, the result is described in terms of membership in fuzzy sets. Defuzzification would transform this result into a single number indicating the trust level of an entity. This may be necessary if we wish to output a real number to the user. An average of maxima method or a centroid method can be used to do this work.

With centroid-based defuzzification, we can get the outcomes V_1=7.2424 ; V_2=7.0424.

It can be seen that the value of factors E_4 is outstandingly low, that means the entity is defective in factors E_4. To get the decider's adequate attention, we should increase its weight to highlight the disadvantages. Variable weight based fuzzy comprehensive evaluation is applied in trust model.

According to variable weight based fuzzy comprehensive evaluation algorithm, we can get the weights of factors and comprehensive trust values on these two sets of evaluations [6, 8, 7, 9] and [7, 9, 8, 4].

Parameters of the first set of evaluations , $\{\lambda_1(u),\lambda_2(u),\lambda_3(u),\lambda_4(u)\}$, can be computed with formula （10）:

$\{\lambda_1(u),\lambda_2(u),\lambda_3(u),\lambda_4(u)\}$= {0.4262, 0.2351, 0.3803, 0.2095}.

Parameters of the second set of evaluations are:

$\{\lambda_1(u),\lambda_2(u),\lambda_3(u),\lambda_4(u)\}$= {0.3805, 0.2119, 0.3432, 0.3900}

Variable weights of every factor on evaluations [6, 8, 7, 9] and [7, 9, 8, 4] are as Tab. Ⅱ:

TABLE II.
VALUES OF VARIABLE WEIGHTS WHEN EVALUATIONS ARE GIVEN

Evaluation	Weight	E_1	E_2	E_3	E_4
[6,8,7,9]	w_i	0.3407	0.1879	0.3040	0.1675
[7,9,8,4]	w_i	0.2870	0.1599	0.2589	0.2942

With variable weight based fuzzy comprehensive evaluation algorithm, the comprehensive trust vectors of valuations[6,8,7,9] and [7,9,8,4] can be gotten:

V_1=[0 0 0 0.4927 0.4236 0.0838];

V_2=[0 0 0.2942 0.1435 0.4824 0.0799];

Defuzzify these two trust vectors by centroid method, we can get the following results:

V_1=7.1347,V_2=6.6500.

Compare this results computed by variable weight based fuzzy comprehensive evaluation and the results computed by fixed weighted fuzzy comprehensive evaluation, we can find that former is always smaller than the latter except that every single factor in the assessment factors have taken the maximum.

The affection of the ill factor have low value can be highlighted by give it a larger weight in variable weight based fuzzy comprehensive evaluation. The comprehensive trust value will then be decreased even if its other factors have high evaluations. By variable weight based fuzzy comprehensive evaluation, the Grid entities can avoid choosing service providers with some defective.

IV. RELIABLE AND LOAD BALANCING REPUTATION SYSTEM

If entity P has direct interacting with entity Q in the past time, it has its own direct trust value on Q. Reputation is an entity's belief in another entity's capabilities, honesty and reliability based on recommendations received from other entities within a specific context at a given time[11] [12]. Reputation is global. Any node in Grid can request and get the same value of another node's reputations. The reputations are distributed stored in the Grid, and can be gotten repeatedly. Reputation system can avoid repeating computing in recommand trust system and can reduce network traffic.

A. *Representation of Reputation*

Reputation of an entity is an expectation of its behavior based on other entities' observations or the collective information about the entity's past behavior within a specific context at a given time.

The latest updating time and the times of trades of the node will affect the reliability and validity of reputation information. Let *n* denotes the times of trades of node *Q* ; *V* denotes the synthetic reputation vector. Reputation of node *Q* in context *c* at time *t* can be described as follows:

$$Q\text{'s } R\text{eputation}_{\ c} \quad value\ \ V \quad T\text{ime t} \quad count\ \ \text{n} \ \cdot$$

B. *Initialization of Reputation*

Because in dynamic grid, nodes and domains can join at any time, proper initialization of reputation is needed.

Identity authentication system can help determine the initial reputation values. In cross-domain trust system, to make new nodes have opportunities to participate in the transactions with nodes in different domains, we set the initial reputation of a new node as T_3 (denotes the "trust" fuzzy set), set the times of trades as 0. Trust level *T3* will give the node opportunities to participate in the transactions, and can avoid high risk from unfamiliar new nodes. The Parameter *n* which denotes the times of trades of node *Q* will give the information that the reliability of the reputation values of *Q*.

C. *Reliable and load balancing storing and accessing mechanisms of reputation system*

In Most literatures, reputations were stored in the node itself or in the domain central node. However, storing in the node itself is not secure enough and Storing in domain central nodes is easy to cause bottleneck.

In this paper, we proposed a Dual Two-Layer Chord Protocol (DTLCP) as reputation storage and retrieval infrastructures. Chord is a distributed lookup protocol proposed by Ion Stoica, etc[14][15]. It is a protocol based on Distributed Hash Tables. By using consistent hash function, a node's identifier is chosen by hashing the node's IP address, while a reputation identifier key is produced by hashing the reputation information. And we use globally participating function that maps reputation keys to node identifiers to insert the information into the network. Reputation key *k* is assigned to the first node whose identifier is equal to or follows (the identifier of) *k* in the identifier space. A Chord node needs only a small amount of "routing" information about other nodes. Because this information is distributed, a node resolves the hash function by communicating with other nodes. Each node uses finger tables[15] to maintain information about other nodes as guide to forward reputation queries.

Chord treated all nodes as the same. If there are more nodes leave or join the Grid, the cost of updating figure tables will increased. However, Participating nodes in Grid are not equivalent. Some super peers, such as domain central nodes or some LAN gateways are more powerful and stable than the other ordinary nodes. The routing information updating cost of network composed of super peers will be much smaller than the entire Grid. And Chord is not reliable for its storing reputation information only in one location.

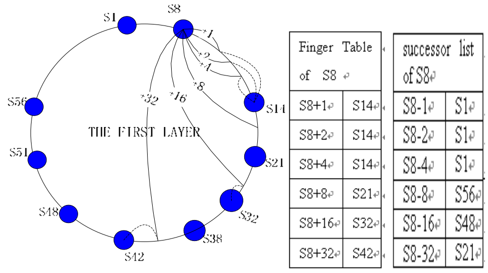
In the Dual two-layer Chord Protocol, we divided the Grid into two layers. The first layer is composed of super peers and the second layer is composed of nodes in one Grid domain.

1) In the first layer, each super node has finger entries at power of two intervals around the identifier circle, each super node can forward a query at least halfway along the remaining distance between the node and the target identifier. Because a node's finger table generally does not contain enough information to directly determine the successor of an arbitrary key *k*, a successor list is stored in each super node. Fig. 3 shows the Chord ring of the first layer and Finger Table and Successor List of super node *S8*.

By hashing the reputation *R*, reputation key of *R* is obtained; and with chord protocol in[18],we can find the first location *S1* of this reputation. Combine this reputation key and identifier of *S1* as identifier *R2*. By hashing *R2*, with the same method, we can find the second location of Reputation *S2*.

2) As the number of nodes in one domain *M* is much smaller than the number *N* in the first layer, in the domains managed by *S1 and S2*, each node has only a successor list, as shown in Fig. 4. Apply consistent hash function in the second layer and find a node to store the reputation information separately; Fig. 4 shows the path taken by a query from node *N6* for reputation key 25 in the second layer of Grid.

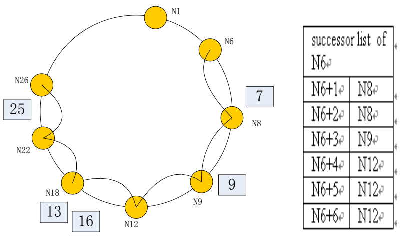
3) When node *P* has a transaction with node *Q*, it will require the reputation of *Q* first. In the support of DTLCP, two nodes with *Q*'s reputations will return reputations of *Q*. If the two values are equal, the reputations are thought trustable. When the transaction is end, update the values in these two locations.



Finger Table of S8	
S8+1	S14
S8+2	S14
S8+4	S14
S8+8	S21
S8+16	S32
S8+32	S42

successor list of S8	
S8-1	S1
S8-2	S1
S8-4	S1
S8-8	S56
S8-16	S48
S8-32	S21

Figure 3. Chord ring of the first layer and Finger Table and successor list of super node S8.



successor list of N6	
N6+1	N8
N6+2	N8
N6+3	N9
N6+4	N12
N6+5	N12
N6+6	N12

Figure 4. The path taken by a query from node *N6* for reputation key 25 in the second layer of Grid

In an N-Domain Grid network, each super node maintains information about only *O(logN)* other super nodes, and a lookup requires *O(logN)* messages. In the support of DTLCP, the reputation system can provide efficient storage and queries that operate in *O(logN)* +*O(M)*overlay hop. *M* is much smaller than *N* in Grid. Chord-based DTLCP can adapt efficiently as nodes join and leave the system, and can answer queries even if the system is continuously changing.

D. *Updating of Reputation*

Regular updating will cause local network congestion. In this paper, we update the reputation by encryption channel immediately after transaction. The Old reputation values have decayed with time. However, the immediate new direct trust value need no time decay. The updating of reputation is described as follows:

$$Q\text{'s Reputation}_{c_1}\ \ value\, V_1\ \ Time\ \mathrm{t}_1\ \ count\ \mathrm{n}$$

$$\wedge P\ \ rectrusts_{c_2}\ Q\ value\ \ V_2\ \ Time\,\mathrm{t}_2$$

$$\wedge\, c_1 \approx c_2$$

$$\Rightarrow Q's\ \mathrm{Re}\,putation\ c_{c_1\circ c_2}\ \ value\,((\gamma(t_1-t_Q,c_1)*V_1)\oplus w*V_2)/(w+*\gamma(t_1-t_Q,c_1))$$
$$\text{Time } \mathrm{t}_2\ \ \text{count } (\mathrm{n}+1)$$

w is weight of the new direct trust value of node *P* on node *Q*. *w* is proportional to the recommend honest reputation of *P* which can be obtained from DTLCP system, and is inversely proportional to the times of transactions of *Q*.

Parameter *count* denotes the times of transactions that node *Q* has participated. It is updated after every transaction and reflects the reliability of the reputation value. For example, a node *A* has transacted 1000 times is more reliable than a node *B* which only transacted 10 times even if their reputation value is the same.

Expression $\gamma(t_1-t_Q,c_1)$ is the time decay function which will affect the weight of old reputation in the updating algorithm.

V. APPLICATION SCENARIOS OF REPUTATION SYSTEM

In a Grid, there are a large number of service providers. Trust and reputation system is needed to help choose more secure and reliable service provides.

Due to the dynamic, open and competitive environment, the application of reputation system in Grid is based on Agent and Virtual Organizations(VO)[13]. The application scenarios can be described as follows:

1) The establishment of virtual organizations: The Agent which wants to get some service initiated service request first; and then resource discovery component localized *n* service providers. A temporary virtual organizations is established, the requesting Agent will be the organization managers of the virtual organization and manager various operations in VO.

2) The manager Agent send bid invitations and service requirements to these *n* service providers. Service providers may be more than one. They can provide a set of services or share a big service task together.

3) At the same time, manager Agent inquires the reputation values of these *n* service providers by DTLCP.

4) Based on the retrieved reputations of these service providers from DTLCP and other feedback information from these service providers, the Agent will decide whether or not to use the services they provided.

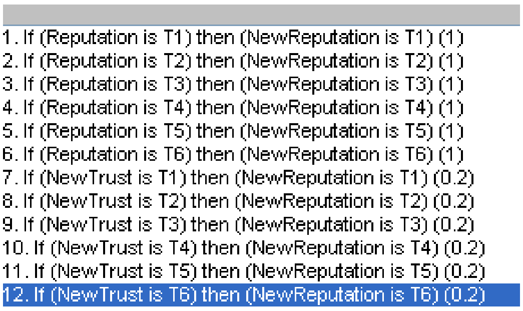
5) During the operation of a VO, the reputations of service providers, such as reputations on implementation of the contract or the quality of services provided are monitored by a Monitor Agent. If there is breach of contract or a drop in the quality of service, the MA will report it to the reputation system, and their reputation will be decreased. A member with too low reputation value will be triggered to exit. In order to ensure the availability of virtual organization, re-formation of VO is needed when necessary.

6) The Services ended and the virtual organizations are revocated.

As can be seen from the above description, the Grid transaction in the support of reputation system is more reliable and more suitable for dynamic, open and complex Grid environment.

VI. SIMULATION EXPERIMENT AND RESULTS ANALYSIS OF FUZZY INFERENCE BASED REPUTATION UPDATE

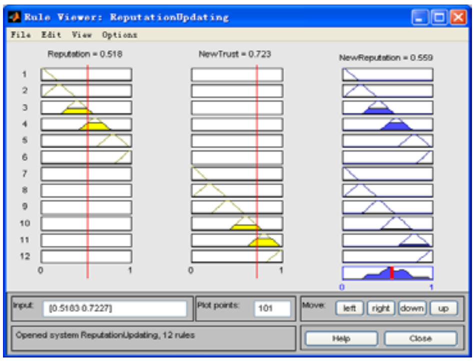
As described in section III, six fuzzy subsets are set to express "reputation" linguistic variables. Suppose weight of the new direct trust *w* is computed as 0.2, and 12 fuzzy reasoning rules are set as shown in Fig. 5. The new updated reputation value is obtained by the aggregation operation of these 12 fuzzy reasoning rules with *fuzzy bounded sum* operator ⊕. Every fuzzy rule has a weight in the aggregation.



1. If (Reputation is T1) then (NewReputation is T1) (1)
2. If (Reputation is T2) then (NewReputation is T2) (1)
3. If (Reputation is T3) then (NewReputation is T3) (1)
4. If (Reputation is T4) then (NewReputation is T4) (1)
5. If (Reputation is T5) then (NewReputation is T5) (1)
6. If (Reputation is T6) then (NewReputation is T6) (1)
7. If (NewTrust is T1) then (NewReputation is T1) (0.2)
8. If (NewTrust is T2) then (NewReputation is T2) (0.2)
9. If (NewTrust is T3) then (NewReputation is T3) (0.2)
10. If (NewTrust is T4) then (NewReputation is T4) (0.2)
11. If (NewTrust is T5) then (NewReputation is T5) (0.2)
12. If (NewTrust is T6) then (NewReputation is T6) (0.2)

Figure 5. Fuzzy Rules of Reputation Updating.

With the original reputation value R_{old} is 0.518, and the new direct trust T_{new} is 0.723, under the support of these fuzzy reasoning rules, we can calculate the new value of the reputation R_{new} =0.559. The rules viewer of all reputation updating fuzzy rules is shown in Fig. 6.



Reputation = 0.518

NewTrust = 0.723

NewReputation = 0.559

Input: [0.5183 0.7227]

Plot points: 101

Opened system ReputationUpdating, 12 rules

Figure 6. Rules viewer of all reputation updating fuzzy rules.

The input and output surface viewer of reputation composing when the weight of new direct trust *w* is 0.2 is shown in Fig. 7.

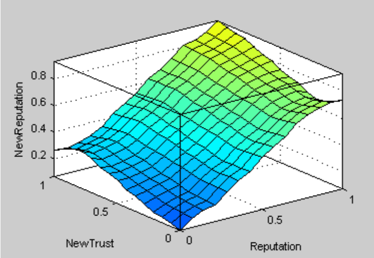


Figure 7. Input and Output Surface Viewer When the Weight of New direct Trust is 0.2.

If the reliability of old reputation of *Q* is not high, that is, the times of transactions of *Q* is small, the last reputation is updated a long time ago or the reputation in recommend context of node *P* transacted with *Q* is relatively high, the weight of new direct trust of *P* on *Q, w,* will be higher. Suppose the weight *w* is computed as 0.6, the input and output surface viewer of reputation composing based on fuzzy reasoning rules is shown in Fig. 8.

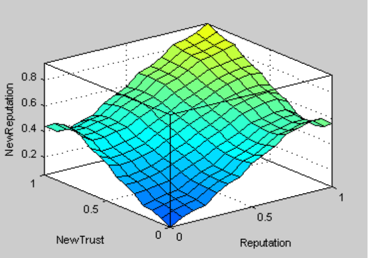


Figure 8. Input and Output Surface Viewer When the Weight of New direct trust is 0.6.

From Fig. 7 and Fig. 8 we can see, when the weight of new direct trust is higher, the new reputation increased with new direct trust faster. That is, if the times of transactions of *Q* is small, the last reputation is updated a long time ago or if the recommend reputation of node *P* is relatively high, the impact of new direct trust on new reputation is higher.

VII. CONCLUSIONS

Trust and Reputation system plays an important role in Grid security field. In this paper, fuzzy theory is used to express and compute trust. The direct trust matrix can be gotten by variable weight based fuzzy comprehensive evaluation. By increasing the weight of the attribute when its value is low, such as the evaluation of factor contract abidance is too small, we can give prominence to deficiency, and then the Grid entities can avoid choosing service providers with some defective.

Reputations are obtained by fuzzy deriving and fuzzy combination. Initialization of reputation and updating of reputation are all discussed in this paper. A DHT based Dual Two-Layer Chord Protocol(DTLCP)is proposed as reputation storage and retrieval infrastructures with which the reputation system can provide efficient storage and queries that operate in $O(logN)$ $+O(M)$overlay hop(N is the number of domains in Grid and M the number of nodes in a domain, $M<<N$). DTLCP can adapt efficiently as nodes join and leave the system, and can answer queries even if the system is continuously changing.

In the support of this reliable fuzzy theory based reputation system, trusts and reputations in Grid can be expressed and computed appropriately and simulation experiment and results analysis demonstrate that entities in Grid can interact with other entities more securely.

In this paper, the fuzzy reasoning rules is established and simplified based on experts experience. In the future further study, we will focus on the automatically generation of fuzzy rules base, with fuzzy neural networks and genetic algorithms.

ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers for their valuable comments and suggestions. This work was supported by the Natural Science Foundation of Jiangsu Province of China under Grant NO.BK2006039.

REFERENCES

[1] Alfarez Abdul-Rahman, Stephen Hailes, "Supporting Trust in Virtual Communities," *33rd Hawaii International Conference on System Sciences(hicss 00)*-Volume 6, 2000 IEEE Press, Dec. 2007, pp. 6007.

[2] Audun Jφsang, "The Beta Reputation System", *15th Bled Electronic Commerce Conference; e-Reality: Constructing the e-Economy; Bled, Slovenia*, June2002.

[3] Andrew whitby, Andun Jφsang; "Filtering Out Unfair Ratings in Bayesian Reputation Systems"; *Icfain Journal of Management Research*. Feb.2005, Vol.IV.No.2 pp.48-64.

[4] Lik Mui, Mojdeh Mohtashemi; "Ratings in Distributed Systems: A Bayesian Approach". *Workshop on Information Technologies and Systems* (WITS'2001).

[5] Tang W, Chen Z, "Research of subjective trust management model based on the fuzzy set theory," *Journal of software*, 2003, 14(8):1401-1408.

[6] Tang W, Hu JB, Chen Z; "Research on a Fuzzy Logic-Based Subjective Trust Management Model," *Journal of Computer Research and Development* 42(10), 2005, pp. 1654~1659.

[7] Shanshan Song, Kai Hwang, and Mikin Macwan, "Fuzzy Trust Integration for Security Enforcement in Grid Computing, " *NPC 2004, LNCS 3222*, pp. 9-21.

[8] Farag Azzedin and Muthucumaru Maheswaran, "Evolving and Managing Trust in Grid Computing Systems," *Proceedings of the 2002 IEEE Canadian Conference on Electrical Computer Engineering*, vol.3, pp. 1424- 1429.

[9] Yao Wang, Julita Vassileva, "Trust and Reputation Model in Peer-to-Peer Networks", *Proceedings of the Third International Conference on Peer-to-Peer Computing* (P2P'03), 2003 IEEE. Sep. 2003, pp.150 – 157.

[10] Peng Zuzeng and Sun Wenyu, *Fuzzy Mathematics and its application*, Wuhan University Press, Sep. 2007.

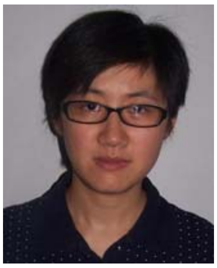
[11] B. K. Alunkal, "Grid Eigen Trust – A Framework for Computing Reputation in Grids," *Master Thesis submitted to Computer Science Department, Illinois Institute of Technology*, Chicago, 2003.

[12] F. Azzedin and M. Maheswaran, "Trust Brokering System and its Application to Resource Management in Public-Resource Grids," *18th International Parallel and Distributed Processing Symposium* (IPDPS'04).Santa Fe: IEEE Computer Society, 2004, pp.22-32.

[13] Jianhua Shao, W Alex Gray, Nick J Fiddian Vikas; "Supporting Formation and Operation of Virtual Organizations in a Grid Environment," *The UK OST e-Science second All Hands Meeting 2004* (AHM'04), Nottingham, UK ,Sep. 2004.

[14] I. Stoica, R. Morris, D.R. Karger, M.F. Kaashoek, H.Balakrishnan, "Chord: a scalable peer-to-peer lookup service for Internet applications", *Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications* (SIGCOMM 2001), ACM Press, 2001, pp. 149–160.

[15] Yuh-Jzer Joung, Jiaw-Chang Wang; "Chord2: A two-layer Chord for reducing maintenance overhead via heterogeneity"; *Computer Networks* 51 (2007) 712–731.



Liao Hongmei was born in Anhui, China on 1977/11/27. She received her B.S. degree in Computer Science and technology from Northeast Normal University(China) in 2000, and her M.S. degree in Computer Science and Technology from China University of Mining and Technology in 2007. Currently, she is a Ph.D. student in School of Computer Science and Technology, China University of Mining and Technology.

She is now a Lecturer in Computer Science and Technology , China University of Mining and Technology. Her research interest is on data fusion, trust model, reputation system, routing protocols in Grid and wireless sensor networks.

Wang Qianping was born in Anhui, China in 1964. He received his Ph.D. degree from Institute of Computing Technology of the Chinese Academy of Sciences in 1997.

In 2000–2002, he visited the Department of Computer Science, Hong Kong University of Science and Technology and researched in "Digital Factory". Now, he is a Professor of the Department of Computer Science and Technology in China University of Mining and Technology. His current research interests include network security and routing protocols in P2P Networks, mobile and pervasive computing and analysis of algorithms for deployment.

Li Guoxin was born in Hebei, China in 1978/10/22. He received his B.S in 2001 from School of Information and Electric Engineering, China University of Mining and Technology.

He is now a Lecturer in Computer Science and Technology , China University of Mining and Technology. He is currently working toward a Ph.D. degree at China University of Mining and Technology. His current research involves security of networks and trust system in P2P and Grid.

Combining Fuzzy Partitions Using Fuzzy Majority Vote and KNN

Chun sheng Li[1] Yaonan Wang[2] Haidong Yang[3]
1 Department of Mathematics and Computational Science, Guang Dong University of Business Studies, Guangzhou, China, 510320, Email: lcs5812084@sina.com
2 College of Electrical and Information Engineering, Hunan University, Changsha, China, 410083
3 College of Automation Science and Engineering, South China University of Technology, Guangzhou, China, 510635

***Abstract*—this paper firstly generalizes majority vote to fuzzy majority vote, then proposes a cluster matching algorithm that is able to establish correspondence among fuzzy clusters from different fuzzy partitions over a common data set. Finally a new combination model of fuzzy partitions is build on the basis of the proposed cluster matching algorithm and fuzzy majority vote. Comparative results show that the proposed combination model is able to foster strengths and circumvent weaknesses of component fuzzy partitions and to combine the component fuzzy partitions into a better fuzzy partition than any of component fuzzy partitions and those resulted from two current combination models of fuzzy partitions.**

***Index Terms*—**fuzzy vote, fuzzy majority vote, combination of fuzzy partitions, evaluation of fuzzy partition

I. INTRODUCTION

Fuzzy clustering has been proved preferable to crisp clustering and a number of fuzzy clustering algorithms [1-4] have been proposed. However, different fuzzy clustering algorithms may produce different fuzzy partitions over the common data set, and none of them are universal enough to perform equally well in any cases. For example, FCM [1] performs well on noiseless dataset with hyper-spherical shape, G-k [2] algorithm on noiseless dataset with hyper-ellipsoidal shape and both AFCM[3] and PFCM[4] are similar to FCM except that they are robust to noises. For a life dataset it may be of different shapes, therefore no single fuzzy clustering algorithm can accurately discover its structure and it makes some errors. However the errors made by different fuzzy clustering algorithms would not necessarily overlap. This suggests that different clusterings potentially offer complementary information about the patterns to be partitioned, which could be harnessed to improve the performance of pattern recognition systems. Therefore, A promising direction for accurate discovery of the data structure may be to combine diverse fuzzy partitions into a consolidate one, which is expected to merge advantages of multiple candidate fuzzy clusterings into one whole. Similar problems associated with crisp clusterings have been studied extensively and there is an extensive body of work on combining multiple crisp clusterings [5-8]. However, the topic of combining fuzzy clusterings has not received the same attention. Evgenia Dimitriadou[9] proposes a combination scheme for fuzzy clusterings that aims to find a consensus fuzzy partition which optimally represents the set of component fuzzy clusterings over the same data set. A.D. Gordon [10] also presents a combination model that aims to identify a consensus fuzzy partition which closely fits the set of component fuzzy partitions over the same data set. However, no theory guarantees that a consensus fuzzy partition representing or fitting a set of fuzzy partitions can represent or fit the real structure of the data set. The current paper also addresses the problem of combining fuzzy partitions with the same number of clusters over the same data set.

There are two difficult problems in combining multiple fuzzy partitions. One is to establish the correspondences among clusters of the component fuzzy partitions so that the first cluster of one partition means the same as that of another one, so is the second cluster and so on, the other problem is to design the rule of combining multiple fuzzy partitions. To solve the first problem, Evgenia Dimitriadou [9] first builds up the confusion matrix between the consensus fuzzy partition that is initialized by one of the component fuzzy partitions and the component fuzzy partition, then the first two clusters associated with the first maximum element of the confusion matrix correspond to each other, so do the second two clusters associated with the second maximum element of the confusion matrix, and so on. Since the initial consensus fuzzy partition is randomly selected from the set of component fuzzy partitions and then updated by each of the other component fuzzy partitions step by step, the resultant consensus fuzzy partition suffers from both the initial consensus fuzzy partition and the order of the component fuzzy partition to take part in updating the consensus fuzzy partition. Unlike Evgenia Dimitriadou[9], A.D. Gordon[10] first builds up the dissimilarity matrix between the consensus fuzzy partition that is initialized randomly and each of the component fuzzy partitions, then treats the problem of cluster correspondence as the problem of assignment and solves it by Hungarian method[11]. The resultant consensus fuzzy partition suffers from the initialization of the consensus fuzzy partition.

To overcome the sensitivity of the above approaches to the initial consensus fuzzy partition in matching clusters from different fuzzy partitions, we transform the

doi:10.4304/jcp.5.5.791-798

problem of establishing the correspondence among the clusters of component fuzzy partitions into the problem of partitioning their cluster centers so that the cluster centers in the same cluster correspond to each other. The second problem is solved by generalizing the majority voting rule for ensemble of crisp partitions to the fuzzy majority voting rule for ensemble of fuzzy partitions. Based on this, a new combination model of fuzzy partitions is build, and its performance are studied intensively by simulation experiments.

The rest of this paper is organized as follows. In section II the related work is reviewed. The traditional majority voting rule is generalized to the fuzzy majority voting rule in section III. An algorithm for matching clusters of different fuzzy partitions is proposed in section IV. A new combination model of fuzzy partitions is proposed in section V .Numerical experiments and conclusions are given in section VI and VII, respectively.

II. Related Work

A The Voting Algorithm [9]

The main idea of literature [9] is to find partition P of a given data set $X=\{x_1, x_1, \cdots, x_N\}$ with g clusters which optimally represents a given set of M partitions of X. Each of these M partitions is represented by an $N\times g$ membership matrix U_h (h=1, 2, …, M). The final partition P_M is encoded as an $N\times g$ matrix. The element $u_{ij}^{(h)}$ of U_h is the degree of membership of x_i to the j-th class of the h-th partition. We denote the i-th row of U_h as $u_i^{(h)}$, that is $u_i^{(h)}$ is the membership vector of the pattern x_i for the partition U_h. The final partition P is encoded as a $N\times g$ matrix with elements p_{ij} and rows p_i. The task of finding an optimal partition is given by the minimization problem:

$$\min_P l(U_1, U_2, \cdots, U_M, P) = \min_{P_1,\cdots,P_N} \min_{\Pi_1,\cdots,\Pi_M} \left(\frac{1}{M} \sum_{h=1}^{M} \frac{1}{N} \sum_{i=1}^{N} \left\| \Pi_h\left(u_i^{(h)}\right) - p_i \right\|^2 \right) \tag{1}$$

Where $\Pi_h(U_h)$ is any permutation of the columns of U_h.

This minimization problem is solved by the voting algorithm[9], which is described in Table I.

Table 1 Voting Algorithm [9]

Step1 set $P^{(1)}=U_1$ and $\hat{\Pi}_1 = id$ (*id* means identical permutation);

Step2 for m=2 to M

(a) compute the solution $\hat{\Pi}_m$ of

$$\max_{\Pi_m} tr\left(\left(\sum_{l=1}^{m-1} \hat{\Pi}_l(U_l)\right)'\right)\Pi_m(U_m) = \max_{\Pi_m} tr\left(\left(P^{(m-1)}\right)'\Pi_m(U_m)\right)$$

by the following approximation algorithm:

(1)build up the confusion matrix between P and U;

(2)find the maximum element in this confusion matrix;

(3)associate the two clusters corresponding to the maximum element;

(4)remove these two clusters;

(5)with the reduced confusion matrix go to (2);

(b) compute the voting result $P^{(m)}$ after m runs as

$$P^{(m)} = \frac{m-1}{m} P^{(m-1)} + \frac{1}{m} \hat{\Pi}_m(U_m)$$

$P^{(m)}$ denotes the voting result after the first m steps.

B A Model for Fitting a Fuzzy Consensus Partition to a Set of Membership Functions [10]

This model identifies the "closest" consensus fuzzy partition P_M fitting its membership function matrix U_M to the membership function matrices $\{U_h$ (h=1,2,…,M)$\}$, that have been permuted to "best" match g classes of P_h with g classes of P_M. The "closest" consensus fuzzy partition P_M of $\{P_h$ (h=1,2,…,M)$\}$ can be obtained by solving the following problem in the integer variables $Y_h=[y_{hpl}]$ and nonnegative membership functions $U_M = \left(u_{ij}^{(M)}\right)$:

$$[\mathrm{P1}]\ \min F(Y_1, Y_2, \cdots, Y_r, U_M) = \sum_{h=1}^{r} w_h \left\|U_h Y_h - U_M\right\|^2 = \sum_{h=1}^{M} \sum_{i=1}^{N} \sum_{p=1}^{g} \sum_{l=1}^{g} w_h \left(u_{ip}^{(h)} - u_{il}^{(M)}\right)^2 y_{hpl} \tag{2}$$

Subject to the constraints

$$\sum_{p=1}^{g} y_{hpl} = 1(l = 1,2,\cdots,g; h = 1,2,\cdots M) \tag{3}$$

$$\sum_{l=1}^{M} y_{hpl} = 1(p = 1,2,\cdots,g; h = 1,2,\cdots M) \tag{4}$$

$$y_{hpl} \in \{0,1\}(p, l = 1,2,\cdots,g; h = 1,2,\cdots M) \tag{5}$$

$$m_{il} \geq 0(i = 1,2,\cdots,N; l = 1,2,\cdots,g) \tag{6}$$

$$\sum_{l=1}^{g} m_{il} = 1(i = 1,2,\cdots,N) \tag{7}$$

The constrained problem [P1] can be minimized by means of the alternating least-square algorithm (*ALS*) described in Table II, that alternates between minimizing F(Y_1, Y_2, $\cdots$, Y_M, U_M) with respect to $\{Y_h$ (h=1,2,…,M)$\}$ given the current estimate of the median membership function matrix U_M; and minimizing F(Y_1, Y_2, $\cdots$, Y_M, U_M) with respect to U_M given the current matching of classes between Y_h and U_M (h=1, 2,…, M).

Table 2 The Alternating Least-square Algorithm (ALS)

Step1 Given the estimates of the median membership function matrix U_M, new least-squares estimates of the elements of Y_h (h=1, 2, …, M) can be determined by solving M independent matching problems:

[P1a] $\min \Delta_1(U_h, U_M, Y_h) = \sum_{i=1}^{N} \sum_{p=1}^{g} \sum_{l=1}^{g} \left(u_{ip}^{(h)} - u_{il}^{(M)}\right) y_{pl}$ $(h = 1,2,\cdots,M)$

Subject to constraints (3), (4) and (5).

([P1a] can be efficiently solved using the well-known Hungarian method [11] in O(g^3) time complexity).

Step2 treating the elements of Y_h (h=1, 2, …, M) as constraints, it is necessary to solve:

[P1b]: min F(Y_1, Y_2, $\cdots$, Y_M, U_M)

Subject to constraints (6) and (7).

The solution is given by

$$u_{il}^{(M)} = \sum_{h=1}^{M} \sum_{p=1}^{g} w_h u_{ip}^{(h)} y_{hpl} \Big/ \sum_{h=1}^{M} w_h \ (i = 1,2,\cdots,N; l = 1,2,\cdots,g)$$

C Majority Voting Rule (MAJ)

This rule does not require the *a posteriori* outputs for each class, and each classification gives only one crisp class output as a vote for that class. Then, the ensemble output is assigned to the class with the maximum number of votes among all classes. For any sample $x \in X$, for a group of M classification in a g-class problem, we denote the decision of label outputs for x from classification $f(i)$ is $c(i)$, $1 \leq c(i) \leq g$. Several terminologies are defined in the following.

Definition 1 For a sample $x \in X$, the crisp vote $d_{i,l}(x)$ for class l given by classification $f(i)$ is defined as

$$d_{il}(x)=\begin{cases}1 & c(i)=l\\ 0 & c(i)\neq l\end{cases} \qquad (8)$$

$d_{i,l}$=1 means that classification *f(i)* votes for class *l*, and $d_{i,l}$=0 means that classification *f(i)* votes against class *l*.

Definition 2 For a sample $x\in X$, the discriminating function $g(l|x)$ for class *l* (1≤*l*≤g) is defined as

$$g(l|x)=\sum_{i=1}^{M}d_{il}(x) \qquad (9)$$

The discriminating function represents the total number of votes given to a class by all classifications. The higher value of the discriminating function $g(l|x)$ indicates more supports for class *l*. Consequently, the output of an ensemble of classifications is the class label with the maximum value of the discriminating function.

$$k=\arg\max_{1\le l\le g} g(l|x) \qquad (10)$$

III Fuzzy Majority Voting Rule

The definition of crisp vote $d_{i,l}(x)$ indicates that a crisp partition *f(i)* either supports or denies utterly that pattern *x* belongs to class *l*. Since a fuzzy partition considers that any patterns belong to all clusters with different membership degrees, the crisp vote has to be revised before it is applied to fuzzy partitions.

Since a crisp clustering is a special case of fuzzy clustering, Definition 1 indicates that the crisp vote $d_{i,l}(x)$ is actually the membership degree of pattern *x* belonging to class *l*. From this point of view the natural way of generalizing a crisp vote to a fuzzy vote is to define the fuzzy vote $\tilde{d}_{il}(x)$ given to pattern *x* by the fuzzy clustering $\tilde{f}(i)$ as the fuzzy membership degree $u_l^{(i)}(x)$ of pattern *x* belonging to class *l* derived from the fuzzy clustering $\tilde{f}(i)$. This yields that the consensus fuzzy partitions is the mean of all the component fuzzy partitions, which is the optimal representation of all the component fuzzy partitions, just as Evgenia Dimitriadou etc al stated in the literature[9]. The reason can be found in remarks at the end of this section. However experiments in section VI show that the mean of all the component fuzzy partitions is not sure to represent the real structure of the data set. Considering this, we do not simply define the fuzzy vote $\tilde{d}_{il}(x)$ as the fuzzy membership degree $u_l^{(i)}(x)$, but treat the classes differently, that is, we directly define $\tilde{d}_{il}(x)=u_l^{(i)}(x)$ for class $l=\arg\max_{1\le t\le g} u_t^{(i)}(x)$, but for other classes, we define

$\tilde{d}_{ik}(x)=u_k^{(i)}(x)\left(1-\max_{1\le t\le g}u_t^{(i)}(x)\right)$, $k\in\{1,2,\cdots,g\}$ and $k\neq l$.

Experiments in section VI show that this definition of fuzzy vote yields relatively good consensus fuzzy partition. The following definitions are the fuzzy counterparts of definitions 1-2.

Definition 3 for a pattern $x\in X$, the fuzzy vote given to class *l* by the fuzzy partition $U^{(i)}=\left(u_l^{(i)}(x)\right)_{N\times g}$ is defined as

$$\tilde{d}_{il}(x)=\begin{cases}u_l^{(i)}(x) & l=\arg\max_{1\le t\le g}u_t^{(i)}(x)\\ u_l^{(i)}(x)\left(1-\max_{1\le t\le g}u_t^{(i)}(x)\right) & otherwise\end{cases} \qquad (11)$$

Where *N* is the number of patterns, *g* the number of clusters and $u_l^{(i)}(x)$ the membership degree of pattern *x* belonging to cluster *l*. Contrary to the crisp vote, the fuzzy vote indicates that a fuzzy partition neither supports nor denies utterly that a pattern belongs to a cluster, but supports it belongs to all clusters to different extents. It is obvious that the crisp vote is the special case of the fuzzy vote.

Definition 4 for a pattern $x\in X$, the fuzzy discriminating function for class *l* (1≤*l*≤*g*) is defined as

$$\tilde{g}(l|x)=\sum_{i=1}^{M}\tilde{d}_{il}(x) \qquad (12)$$

Where *M* is the number of component fuzzy partitions. Like the discriminating function defined by formula (9), the fuzzy discriminating function of a class also represents the amount of supports given to it by all fuzzy partitions. Higher value of the fuzzy discriminating function $\tilde{g}(l|x)$ means more supports for pattern *x* belonging to class *l*. Unlike simple majority voting rule, instead of assigning a class label to pattern *x*, we calculate the membership degree $cu_l(x)$ of *x* belonging to each class *l*, 1≤*l*≤*g*, determined by *M* fuzzy partitions jointly as follows

$$cu_l(x)=\tilde{g}(l|x)\Big/\sum_{t=1}^{g}\tilde{g}(t|x), 1\le l\le g \qquad (13)$$

Formula (13) indicates that the combination of fuzzy partitions is still a fuzzy partition. This is different from the consensus crisp partition. If formula (13) is replaced with $k=\arg\max_{1\le l\le g}\tilde{g}(l|x)$, it is obvious that the majority voting rule is the special case of the fuzzy majority voting rule.

In the following we exemplify the fuzzy majority voting rule. Supposing that there are three component fuzzy partitions over the same data set, each of which has three clusters and is denoted by its fuzzy partition matrix $U^{(i)}$ (*i*=1, 2, 3). In the case that the correspondence among clusters from the component fuzzy partitions is established, i.e., the first column of the fuzzy component partitions represents the same class, so are the second and third column. Given a pattern *x*, the membership degree of *x* belonging to each cluster derived from $U^{(i)}$ (*i*=1,2,3) is

$$\left(u_1^{(1)}(x),u_2^{(1)}(x),u_3^{(1)}(x)\right)=(0.1894,0.6894,0.1212)$$

$$\left(u_1^{(2)}(x),u_2^{(2)}(x),u_3^{(2)}(x)\right)=(0.4187,0.2761,\ 0.3052)$$

$$\left(u_1^{(3)}(x),u_2^{(3)}(x),u_3^{(3)}(x)\right)=(0.2527,0.4743,0.2730)$$

Definition 3 yields

$$\left(\tilde{d}_{11}(x),\tilde{d}_{12}(x),\tilde{d}_{13}(x)\right)=(0.0588,0.6894,0.0376).$$

$$\left(\tilde{d}_{21}(x),\tilde{d}_{22}(x),\tilde{d}_{23}(x)\right)=(0.4187, 0.1605, 0.1774)$$
$$\left(\tilde{d}_{31}(x),\tilde{d}_{32}(x),\tilde{d}_{33}(x)\right)=(0.1328,0.4743,0.1435)$$

Formular (12) gives

$$\left(\tilde{g}(1|x),\tilde{g}(2|x),\tilde{g}(3|x)\right)=(0.6103, 1.3242, 0.3585)$$

Formula (13) results in
$(cu_1(x), cu_2(x), cu_3(x))$=(0.2662, 0.5775, 0.1563) .

Remarks: if $\tilde{d}_{il}(x)=u_{il}(x)$, then

$$\begin{aligned} cu_l(x) &= \frac{\tilde{g}(l|x)}{\sum_{t=1}^{g}\tilde{g}(t|x)} \\ &= \frac{\sum_{i=1}^{M}u_l^{(i)}(x)}{\sum_{t=1}^{g}\sum_{i=1}^{M}u_t^{(i)}(x)} \\ &= \frac{\sum_{i=1}^{M}u_l^{(i)}(x)}{\sum_{i=1}^{M}\sum_{t=1}^{g}u_t^{(i)}(x)} \\ &= \frac{1}{M}\sum_{i=1}^{M}u_l^{(i)}(x) \end{aligned}$$

IV A Cluster Matching Algorithm Based on KNN

When no a priori class information for the patterns is available, a direct application of fuzzy majority voting rule to combining fuzzy partitions is not possible, for it is not immediately clear which cluster from a specific partition corresponds to what in another. Therefore, it is necessary to establish the correspondence among clusters of all component fuzzy partitions so that the same column of $U^{(i)}=\left(u_l^{(i)}(x)\right)_{N\times g}, i=1,2,\cdots,M$ defines the same cluster. Both Evgenia Dimitriadou[9] and A.D. GORDON[10] establish this kind of correspondence by corresponding the clusters of each component fuzzy partition with those of the consensus fuzzy partition, which is firstly initialized randomly or with one of the component fuzzy partitions, then updated adaptively. They suffer from the initialization of the consensus fuzzy partition.The underlying idea of establishing the correspondence between clusters of one fuzzy partition and those of another is that the similar clusters correspond to each other so that the sum of dissimilarities between two fuzzy partitions is minimized, as shown in the literatures [9, 10]. Inspired by this idea, we transfer the problem of pairing clusters from different fuzzy partitions into the problem of partitioning the set of cluster centers. Supposing that there are M fuzzy partitions $U^{(1)},U^{(2)},\cdots,U^{(M)}$, each of which has g clusters. Each cluster is represented by its center. This yields a set of cluster centers, $V=\left\{v_1^{(1)},\cdots,v_g^{(1)},v_1^{(2)},\cdots,v_g^{(2)},\cdots,v_1^{(M)},\cdots,v_g^{(M)}\right\}$.where $v_j^{(i)}$ is the j-th cluster center of the i-th fuzzy partition. We define the similarity between two clusters as the Euclidian distance between their centers. Consequently, establishing the correspondence among the clusters of M fuzzy partitions is transferred into partitioning the set V of center vectors into g clusters, each of which contains M center vectors from different fuzzy partitions. The center vectors belonging to the same cluster correspond to each other, i.e., if $v_{j_1}^{(1)},v_{j_2}^{(2)},\cdots,v_{j_M}^{(M)}$ belong to the same cluster, then the j_1-th cluster of $U^{(1)}$, the j_2-th cluster of $U^{(2)}$, …, the j_M-*th* cluster of $U^{(M)}$ define the same cluster, where $s\in\{1,2,\cdots,M\}, j_s\in\{1,2,\cdots,g\}$. To assure the center vectors in the same cluster are from different fuzzy partitions, we define the dissimilarity between two cluster centers as

$$d\left(v_j^{(i)},v_t^{(s)}\right)=\begin{cases}\left\|v_j^{(i)}-v_t^{(s)}\right\|_2^2, & i\neq s\\ +\infty\quad, & i=s, j\neq t\end{cases}\tag{14}$$

The K nearest neighbors method (KNN) is employed to partition the data set V into g clusters, each of which contains M center vectors. Consequently, an approach to establishing the correspondence among the clusters from different fuzzy partitions is developed, which is described by the pseudo code in Table 3.

Table 3 The Cluster-matching Algorithm Based on KNN

```
1 Compute the centre vectors of each fuzzy partition by
    $v_k^{(i)}=\sum_{x\in X}u_k^{(i)}(x)\cdot x/\sum_{x\in X}u_k^{(i)}(x), k=1,\cdots,g;i=1,\cdots,M$
2Compute dissimilarity $d\left(v_j^{(i)},v_t^{(s)}\right), j,t=1,\cdots,g$ ,
    $i,s=1,\cdots,M$ using (14);
3Find $M$ nearest neighbours for each centre vector $v_t^{(i)},t=1,\cdots,g;i=1,\cdots,M$ . They form a $M$-nearest neighbourhood denoted by $\{v_t^{(i)}\}$;
4Compute the compactness $comp(\cdot)$ of each $M$-nearest neighbourhood by $comp\left(\{v_t^{(i)}\}\right)=\sum_{v_1,v_2\in\{v_t^{(i)}\}}d(v_1,v_2)$;
5Select $g$ most compact and disjoint $M$-nearest neighbourhoods. The centre vectors belonging to the same $M$-nearest neighbourhood represent the same cluster and the columns corresponding to them in $U^{(i)}$ ($i$=1,2, …,$M$) are labeled as the same class label.
```

In the following we exemplify the proposed cluster-matching algorithm. Supposing that there are three component fuzzy partitions, each of which has three clusters. Their cluster centers are listed in Table 4. The dissimilarity between any pair of cluster centers is derived from formula (14). The M-nearest neighbourhood of each center vector and its compactness are listed in Table 4, where M=3. The set of nine cluster center vectors is partitioned into three disjoint clusters: $\left\{v_1^{(1)},v_2^{(2)},v_2^{(3)}\right\}$, $\left\{v_2^{(1)},v_1^{(2)},v_1^{(3)}\right\}$ and $\left\{v_3^{(1)},v_3^{(2)},v_3^{(3)}\right\}$. The cluster $\left\{v_1^{(1)},v_2^{(2)},v_2^{(3)}\right\}$ means that the first cluster of the first fuzzy partition, the second cluster of the second partition and the second cluster of the third fuzzy partition define the same cluster, so do $\left\{v_2^{(1)},v_1^{(2)},v_1^{(3)}\right\}$ and $\left\{v_3^{(1)},v_3^{(2)},v_3^{(3)}\right\}$.

V A Combination Scheme for Fuzzy Partitions Using Fuzzy Majority Voting Rule and KNN

Supposing that there are M fuzzy partitions over the data set X, each of which is denoted by a fuzzy partition matrix $U^{(i)}$, i=1, 2, …, M. They are matched by the cluster matching algorithm in Table 3, then the well matched *M* fuzzy partitions are combined into a consensus fuzzy partition using fuzzy majority voting rule. The pseudo description of the proposed combination model is given in Table 5.

Table 3 The Cluster Centers of Three Component Fuzzy Partitions

$v_1^{(1)}$	$v_2^{(1)}$	$v_3^{(1)}$	$v_1^{(2)}$	$v_2^{(2)}$	$v_3^{(2)}$	$v_1^{(3)}$	$v_2^{(3)}$	$v_3^{(3)}$
6.6294	5.0626	5.9525	5.0626	6.6294	5.9525	5.0626	6.6294	5.9525
3.0091	3.3618	2.7947	3.3618	3.0091	2.7947	3.3618	3.0091	2.7947
5.4126	1.6670	4.4202	1.6670	5.4126	4.4202	1.6670	5.4126	4.4202
1.9359	0.3271	1.4351	0.3271	1.9359	1.4351	0.3271	1.9359	1.4351

Table 4 the 3-nearest neighbourhood of each cluster center and its compactness

	3-nearest neighbourhood	compactness
$v_1^{(1)}$	$\{v_1^{(1)}, v_2^{(2)}, v_2^{(3)}\}$	0.2287*10^{-5}
$v_2^{(1)}$	$\{v_2^{(1)}, v_1^{(2)}, v_1^{(3)}\}$	0.0408*10^{-5}
$v_3^{(1)}$	$\{v_3^{(1)}, v_3^{(2)}, v_3^{(3)}\}$	0.2593*10^{-5}
$v_1^{(2)}$	$\{v_1^{(2)}, v_1^{(3)}, v_2^{(1)}\}$	0.0408*10^{-5}
$v_2^{(2)}$	$\{v_2^{(2)}, v_2^{(3)}, v_1^{(1)}\}$	0.2287*10^{-5}
$v_3^{(2)}$	$\{v_3^{(2)}, v_3^{(3)}, v_3^{(1)}\}$	0.2593*10^{-5}
$v_1^{(3)}$	$\{v_1^{(3)}, v_1^{(2)}, v_2^{(1)}\}$	0.0408*10^{-5}
$v_2^{(3)}$	$\{v_2^{(3)}, v_2^{(2)}, v_1^{(1)}\}$	0.2287*10^{-5}
$v_3^{(3)}$	$\{v_3^{(3)}, v_3^{(2)}, v_3^{(1)}\}$	0.2593*10^{-5}

Table 5 A Combination Model of Fuzzy Partitions Based on Fuzzy Majority Voting Rule and KNN

Step1 input the fuzzy partition matrices $U^{(i)}$ (*i*=1, 2, …, *M)* and pattern samples *X*;

Step2 establish the correspondence among clusters from *M* fuzzy partitions using the cluster-matching algorithm in Table 3;

Step3 combine *M* fuzzy partitions using fuzzy majority voting rule described in section III;

V Experiments

An important consideration in the combination of partitions is that much better results can be achieved if diverse partitions, rather than similar partitions, are combined. To create diverse fuzzy partitions we employ three fuzzy clustering algorithms—FCM[1], PFCM[2] and AFCM[3], each of which (except PFCM[2] that is initialized by the output of FCM) is initialized by three centre initialization methods—CCIA[12], kd-tree[13] and MST[14], respectively. Therefore, there are totally nine fuzzy partitions denoted by FCM-CCIA, FCM-MST, FCM-kd-tree, PFCM-CCIA, PFCM-MST, PFCM-kd-tree, AFCM-CCIA, AFCM-MST, AFCM-kd-tree, respectively. They are combined into a consensus fuzzy partition, denoted by FMV, by the combination model in Table 5, in the way depicted in Fig. 1.

To test the performance of the proposed combination model, we compare it with two combination methods *voting* [9] and *ALS*[10] on four real data sets, which are described in Table 6.

For all fuzzy clustering algorithms we use the following *Computational Protocols*; convergence term ε=0.0001, maximum number of iterations=100, the fuzzifer *m*=2. The parameters of PFCM are initialized as follows: *m*=2, *η*=1.5, *a*=1, *b*=3. *ALS* suffers from the initialization of the consensus fuzzy partition. We initialize it with each of nine component fuzzy partitions respectively. The *voting* [9] algorithm suffers from the sequence of the component fuzzy partitions to take part in the combination of fuzzy partitions. We place nine component fuzzy partitions in the order of FCM-CCIA, FCM-MST, FCM-kd-tree, AFCM-CCIA, AFCM-MST, AFCM-kd-tree, PFCM-CCIA, PFCM-MST, PFCM-kd-tree, then initialize the consensus fuzzy partition $P^{(1)}$ with each of the above nine fuzzy partitions, respectively and fix others in their places.

We evaluate the fuzzy partition using pattern recognition rate PR that is a standard evaluation index, partition coefficient PC[18] that measures the fuzzy degree of fuzzy partitions, fuzzy Rand indexe ω_R and related indexes－fuzzy Jaccard coefficient ω_{JC}, fuzzy Fowlkes-Mallows index ω_{FM}, fuzzy Minkowski measure ω_M and fuzzy Γ statistic ω_Γ [17], which are objective criteria for the evaluation of fuzzy partitions, as R. J. G. B. Campello stated [17]. The big values of the indexes ω_R, ω_{JC}, ω_{FM} and ω_Γ indicate the good

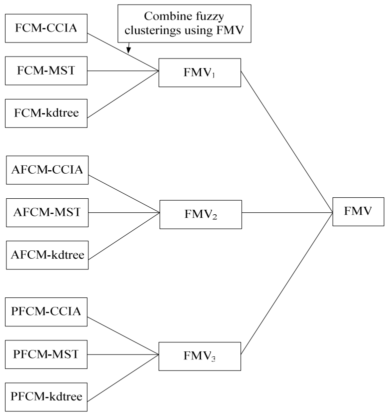


Fig.1 the flowchart of combining fuzzy partitions

(FMV$_i$ (i=1, 2, 3) means the combination results of the first time and FMV the final result of combination.)

Table 6 The brief description of data sets

	Number of instances	Number of attributes	Number of clusters
Pima-Indians-diabete [15]	768	8	2
*svmguide*3[16]	1243	22	2
ionosphere[16]	351	34	2
Sat.image[16]	4435	36	6

closeness between the reference partition and the fuzzy partition to be evaluated, while the low value of ω_M reveals good closeness between them.

To test the performance of the proposed combination model of fuzzy partitions, we compare it with all component fuzzy partitions and two consensus fuzzy partitions, *voting* [9] and *ALS*[10], over a small size and a large size data set, ionosphere [15] and sat.image [16]. We also compare it with only two consensus fuzzy partitions, *voting* [9] and *ALS* [10], over two middle size data sets, *diabetes* and *svmguide3*. The comparative results listed in Table 7-9 indicate that, in terms of all evaluation indexes for fuzzy partitions, the proposed combination model FMV outperforms two consensus fuzzy partitions, *voting* [9] and *ALS* [10]. Table 7 and 9 show that, in terms of pattern recognition rate, the proposed combination model FMV is comparable to the best component fuzzy partitions FCM-CCIA, FCM-MST and FCM-kdtree, while in terms of other indexes, FMV is preferable to them over the data sets ionosphere and sat. image, respectively. Table 10 indicates that, in terms of CPU time, *voting* [9] is the cheapest, and FMV is a little more expensive computation than *ALS* [10]. In a word, the proposed combination method FMV is able to foster strengths and circumvent weaknesses of component fuzzy partitions and outperforms *voting* [9] and *ALS* [10] at the cost of a little extra computation in our experiments. It is important for the consensus fuzzy partition to be comparable, but uncertain to be preferable, to the best component fuzzy partition in any cases, for no fuzzy clustering algorithm can generates good partitions in all cases and we do not know which fuzzy clustering algorithm may produce a good clustering in advance. Furthermore, we do not know how to accurately assess a fuzzy partition, much less select the best individual fuzzy partition when no information about the data set is available. In this sense, the consensus fuzzy partition obtained from the proposed combination model is more stable and reliable than any component fuzzy partition, for it can combine the advantages of all component fuzzy partitions and pool them into a consensus fuzzy partition that is uncertain to be better than any component fuzzy partition, but sure to be superior to an overwhelming majority of the component fuzzy partitions in any cases.

To compare FMV with *voting* [9], *ALS* [10] and component fuzzy partitions, we test them over the data set ionosphere. Table 7 shows that all evaluation indexes agree with our consensus fuzzy partition FMV outperforms *ALS* and *voting* a little. Compared to nine individual fuzzy partitions, in terms of pattern recognition rate, the consensus fuzzy partition generated by the proposed combination model is as good as the best component fuzzy partition FCM-CCIA, FCM-MST and FCM-kdtree, better than other six component fuzzy partitions. Table 7 also shows that other evaluation indexes agree with that FMV outperforms all the component fuzzy partitions. In general, the proposed consensus fuzzy partition FMV is able to combine the advantages of all component fuzzy partitions and at least comparable to the best component fuzzy partition. Contrary to FMV, in terms of pattern recognition rate, *ALS* and *voting* are worse than the best component fuzzy partition. Furthermore, *ALS* and *voting* are worse than part of the component fuzzy partitions in terms of other evaluation indexes. In a word, *ALS* and *voting* fail to combine the advantages of the component fuzzy partitions and are inferior to the best component fuzzy partitions. It is also revealed by Table 7 that when the pattern recognition rates of FMV, FCM-MST and FCM-kdtree are equal, other indexes for fuzzy partitions still can distinguish the three fuzzy partitions. This indicates that it is not sufficient to evaluate fuzzy partitions only by pattern recognition rate and other evaluation indexes for fuzzy partitions are also powerful tools for assessing fuzzy partitions.

To further compare the performances of FMV, *ALS* and *voting*, we test them over two middle size data sets, *diabetes* and *svmguide*3. Table 8 shows that FMV outperforms *ALS* and *voting* by huge margins in terms of pattern recognition. The partition coefficients PC listed in Table 8 show that *ALS* and *voting* average conflicting memberships towards 1/*c* (*c*=2) over the data set *diabetes*. The fuzzy rand index and related indexes listed in Table 8 agree with that FMV is a little better than *ALS* and *voting*. Table 8 also shows that *ALS* is as bad as *voting* in terms of all evaluation indexes.

Table 7 the comparative results among consensus fuzzy partitions and individual fuzzy partitions over *ionosphere*

Fuzzy partitions		PR(%)	PC	ω_R	ω_{JC}	ω_{FM}	ω_Γ	ω_M
FMV		70.9402	0.8151	0.5602	0.4056	0.5775	0.1205	0.9038
ALS	Mean	70.3704	0.5303	0.5161	0.3671	0.5373	0.0312	0.9480
	std	0	0	0.0000	0	0	0	0
voting	Mean	70.3704	0.5303	0.5161	0.3671	0.5373	0.0312	0.9480
	std	0	0.0000	0	0	0	0.0000	0
FCM-CCIA		70.9402	0.6487	0.5364	0.3899	0.5612	0.0703	0.9279
FCM-MST		70.9402	0.6487	0.5364	0.3899	0.5612	0.0703	0.9279
FCM-kdtree		70.9402	0.6487	0.5364	0.3899	0.5612	0.0703	0.9279
AFCM-CCIA		69.5157	0.5000	0.5000	0.3500	0.5189	0.0000	0.9636
AFCM-MST		69.5157	0.5000	0.5000	0.3500	0.5189	0.0000	0.9636
AFCM-kdtree		69.5157	0.5000	0.5000	0.3500	0.5189	0.0000	0.9636
PFCM-CCIA		69.2308	0.6423	0.5338	0.3847	0.5558	0.0662	0.9305
PFCM-MST		53.8462	0.5000	0.5000	0.3500	0.5189	0.0000	0.9636
PFCM-kdtree		60.3989	0.5031	0.5015	0.3518	0.5208	0.0028	0.9622

The bold number means the optimal value of evaluation index

Table 8 the comparative results among consensus fuzzy partitions over two small size data sets with two clusters

Data set	Consensus fuzzy partitions		PR(%)	PC	ω_R	ω_{JC}	ω_{FM}	ω_Γ	ω_M
diabetes	FMV		**65.8854**	**0.7766**	**0.5252**	**0.4141**	**0.5863**	**0.0330**	**0.9334**
	ALS	Mean	61.9792	0.5990	0.5118	0.3771	0.5477	0.0174	0.9465
		std	0	0	0	0	0	0	0
	voting	Mean	61.9792	0.5990	0.5118	0.3771	0.5477	0.0174	0.9465
		std	0	0.0000	0.0000	0.0000	0	0.0000	0.0000
*svmguide*3	FMV		**57.2808**	**0.7696**	**0.5076**	**0.3980**	**0.5731**	**0.0121**	**0.8793**
	ALS	Mean	50.6034	0.5406	0.5061	0.3975	0.5725	0.0077	0.8806
		std	0	0	0	0.0000	0	0	0.0000
	voting	Mean	50.6034	0.5406	0.5061	0.3975	0.5725	0.0077	0.8806
		std	0	0.0000	0	0.0000	0	0	0.0000

To further investigate the performance of FMV, we compare it with two consensus fuzzy partitions, *voting* [9] and *ALS* [10], and all the component fuzzy partitions over a large size data set with six clusters. Table 9 shows that three consensus fuzzy partitions, FMV, *voting* [9] and *ALS* [10], and three individual fuzzy partitions, FCM-CCIA, FCM-MST and FCM-kdtree, are very close to each other and FMV is a little better than the best individual fuzzy partition in terms of pattern recognition rate. In terms of other evaluation indexes, *voting* [9] and *ALS* [10] are inferior to FCM-CCIA, FCM-MST, FCM-kdtree, PFCM-CCIA and PFCM-MST, while FMV is superior to all component individual fuzzy partitions, as shown by Table 9. This experiment once more confirms that FMV is able to combine the advantages of component individual fuzzy partitions and at least comparable to the best component fuzzy partition in the case of large size data sets with multiple clusters, while *voting* [9] and *ALS* [10] can not.

Table 9 the comparative results among consensus fuzzy partitions and individual fuzzy partitions over Sat. image

Fuzzy partitions		PR(%)	PC	ω_R	ω_{JC}	ω_{FM}	ω_Γ	ω_M
FMV		**70.9583**	**0.6067**	**0.7883**	**0.3959**	**0.5737**	**0.4895**	**1.0090**
ALS	Mean	70.5850	0.3700	0.6955	0.2974	0.4732	0.3814	1.1995
	std	0.4152	0.0003	0.0010	0.0013	0.0014	0.0019	0.0032
Voting	Mean	69.8409	0.3657	0.6950	0.2959	0.4713	0.3793	1.2009
	std	0.8031	0.0034	0.0027	0.0028	0.0034	0.0044	0.0038
FCM-CCIA		70.3495	0.4799	0.7407	0.3262	0.4995	0.4034	1.1090
FCM-MST		70.3495	0.4799	0.7407	0.3262	0.4995	0.4034	1.1090
FCM-kdtree		70.3495	0.4799	0.7407	0.3262	0.4995	0.4034	1.1090
AFCM-CCIA		59.6843	0.3570	0.6661	0.2552	0.4174	0.3198	1.2386
AFCM-MST		63.7880	0.3675	0.6761	0.3056	0.4934	0.4082	1.2467
AFCM-kdtree		60.2029	0.3340	0.6382	0.2389	0.3985	0.2988	1.2816
PFCM-CCIA		62.6607	0.4223	0.7083	0.3253	0.5027	0.4159	1.1388
PFCM-MST		66.8997	0.4383	0.7198	0.3238	0.5031	0.4165	1.1538
PFCM-kdtree		53.7317	0.3685	0.6543	0.2781	0.4515	0.3542	1.2412

The performance of an algorithm is one important aspect and the computational complexity is another important aspect of the algorithm. So the computational complexities of the proposed combination model FMV, Voting[9] and ALS[10] are also compared in terms of CPU times. The comparative results listed in Table 10 show that Voting[9] is of the cheapest computation and FMV is a little more expensive computation than *ALS*[10].

Table 10 Computation Complexity of Three Combination Methods

Data set	CPU time (second)		
	FMV	ALS	*voting*
ionosphere	0.110129	0.082110	0.026875
svmguide3	0.299652	0.113652	0.067559
diabetes	0.167901	0.137785	0.045441
Sat. image	3.086964	2.117051	0.790977

The computer system is of Genuine Intel ® CPU 2140, 1.60GHz and 1.60GHz, 1GB memory. CPU time is only composed of the running time of FMV, ALS and *voting*, but not that of any individual fuzzy clustering algorithm.

VII CONCLUSIONS

This paper generalizes the traditional majority voting rule to the fuzzy majority voting rule and proposes a cluster matching algorithm, based on which a combination model of fuzzy partitions is developed. We compare our combination method with other two combination methods — *voting* [9] and *ALS* [10] and individual fuzzy partitions. Comparative results show that our combination method outperforms *voting* [9] and *ALS* [10] in terms of all evaluation indexes used in this paper. The reason may be that both *voting* [9] and *ALS* [10] aim to find the consensus fuzzy partition that optimally represents and closely fits the set of component fuzzy partitions, and the optimal representation and fitting of a set of fuzzy partitions do not equal the optimal representation of the real structure of the data set, that is, if the consensus fuzzy partition optimally represents or fits a collection of fuzzy partitions, it does not guarantee to represent the real structure of the data set. We also find that *voting* [9] and *ALS* [10] are a little worse than some of the component individual fuzzy partitions, while FMV is at least comparable to the best one of the component individual fuzzy partitions in all cases. This confirms that FMV is able to foster strengths and circumvent weaknesses of component fuzzy partitions, while *voting* [9] and ALS [10] can not. In a word, FMV is not only superior to *voting* [9] and *ALS* [10], but also more stable and reliable than any individual fuzzy partition in some cases.

It is still important for the consensus fuzzy partition to be comparable to, but not sure to outperform, the best component fuzzy partition in any cases, for no fuzzy clustering algorithm can generates good partitions in all cases and we do not know which fuzzy clustering algorithm may produce a good clustering over a given data set in advance. Furthermore, when no information about the data set is available, it is hard to us accurately evaluate the fuzzy partition, much less pick out the best individual fuzzy partition. In this sense, the consensus fuzzy partition is more stable and reliable than any component individual fuzzy partition, for it is able to combine multiple fuzzy partitions into a consolidate one that is at least comparable to the best component fuzzy partitions in any cases.

Acknowledgements

The project is supported by the National Natural Science Foundation under Grant No. 60835004 and the Scientific Research Foundation for Doctor Grant No. 8451064101000630.

RERENCES

[1] C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, New York: Plenum Press, 1981.

[2] D. E. Gustafson, W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in *Proc. IEEE Conf. Decision Contr.*, San Diego, CA, pp. 761–766, 1979 .

[3] Kuo-Lung Wu, Miin-Shen Yang, "Alternative C-means clustering algorithm," Pattern Recognition, vol.35, , pp.2267-2278, 2002.

[4] Nikhil R. Pal, Kuhu Pal, James M. Keller, James C. Bezdek, "a possibilstic fuzzy c-means clustering algorithm, " IEEE Trans. On Fuzzy systems, vol.13(4), pp517-530, 2005.

[5] Ana L.N. Fred, Anil K. Jain, "Combining Multiple Clusterings Using Evidence Accumulation," IEEE Transactions on pattern analysis and machine intelligence, Vol. 27(6), pp835 – 850, 2005.

[6] Alexander Topchy, Anil K. Jain and William Punch, "Clustering ensembles: models of consensus and weak partitions, " IEEE Transactions on pattern analysis and machine intelligence, VOL. 27(12), pp1866–1881, 2005.

[7] Tianming Hu, Ying Yu, Jinzhi Xiong and Sam Yuan Sung, "Maximum likelihood combination of multiple clusterings," Pattern Recognition Letters 27, pp.1457–1464, 2006.

[8] Alexander Topchy, Anil K. Jain, and William Punch, "Combining Multiple Weak Clusterings," Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03).

[9] Evgenia Dimitriadou, Andreas Weingessel, and Kurt Hornik, "A combination scheme for fuzzy clustering," International Journal of Pattern Recognition and Artificial Intelligence, Vol.16(7), pp.901-912, 2002.

[10] A.D. GORDON, M. VICHI, "Fuzzy partition models for fitting a set of partitions," Psychometrika, VOL. 66(2), pp.229—248, 2002.

[11]G. Carpento, S. Maxtello, E. Toth, "Algorithms and codes for the assignment problem," In B. Simeone, E Toth, G. Gallo, E Maffioli, & S. Pallottino (Eds.), Fortan codes for network optimization. Annals of Operations Research, 13, pp.193-224, 1988.

[12]S. Shehroz Khan, Amir Ahmad, "cluster center initialization algorithm for K-means clustering, " Pattern Recognition Letters 25, pp.1293-1302, 2004.

[13]S.J.,Redmond, C.,Heneghan,"A method for initializing the K-means clustering algorithm using kd-trees,"
Pattern Recognition Letters, vol.28, pp 965-973, 2007.

[14] Yao-nan Wang, Chun-sheng Li, Yi Zuo, "A Selection Model for Optimal Fuzzy Clustering Algorithm and Number of Clusters Based on Competitive Comprehensive Fuzzy Evaluation," IEEE Transaction On Fuzzy Systems, vol 17(3), pp. 568-577, June 2009.

[15] The UCI Machine Learning Repository, 1993, http://www.ics.uci.edu/~mlearn/MLRepository.html

[16] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[17] R.J.G.B Campello, "A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment," Pattern Recognition Letters 28, pp.833-841, 2007.

[18] E.Trauwaert, , "On the meaning of Dunn's partition coefficient for fuzzy clusters," Fuzzy Sets Systems 25, pp.217–242, 1988.

Li Chun-sheng received the B.S and the Master degree in Mathematic from Jiang Xi Normal University in 1992 and the Central South University in 1999 respectively, China. He now is a Ph.D. at the College of Electrical and Information Engineering, Hunan University. His research interests are computational intelligence, intelligent control and intelligent information processing.

Factors Influencing the Adoption of Mobile Service in China: An Integration of TAM

Sun Quan
Business Department, Suzhou Vocational University, Suzhou, China
Email: sunxjtu@163.com

Cao Hao[1] and You Jianxin
School of Economics and Management, Chinese Academy of S& T Management, Tongji University, Shanghai, China
Email: caohao309@163.com

Abstract—**This study based on the technology acceptance model(TAM) and theory of planned behaviour (TPB) presents an extended TAM that integrates trust related construct ('perceived credibility') and resource-related constructs ('perceived cost') into the TAM to analyzing adoption behavior of mobile service. The proposed model was empirically tested using data collected from a survey of mobile service consumers. The structural equation modeling technique was used to evaluate the causal model and confirmatory factor analysis was performed to examine the reliability and validity of the measurement model. Our findings indicated that all variables except perceived ease of use significantly affected users' behavioral intention. The implication of this work to both researchers and practitioners is discussed.**

Index Terms—**Mobile services; TAM; Perceived credibility; Perceived behavioral control**

I. INTRODUCTION

Wireless Internet is the hot topic and trend for both telecom industry and Internet industry, and many dealers now plan to invest greatly for the development of related software and services. Mobile service means that by using mobile terminal equipment, consumers may conduct a vast area of activity comprised of transactions of services, goods and information with monetary value via wireless network. With the growth of technical development, diffusion of mobile device and the higher acceptance consumers have toward mobile service, the domestic commercial market will become larger gradually every year. And with the incorporation of Java mobile phone and 3G systems, the benefit brought by the development, therefore, arouses the attention on industry, government and school. At the beginning for the business, to understand the need and acceptance consumers have toward the application of mobile commerce would be helpful for dealers to evaluate the direction for future development.

This issue has drawn a lot of attention from researchers to understand the factors that drive individuals' adoption/rejection of this innovation. Many studies have been conducted using traditional adoption models and theories such as the Technology Acceptance Model (TAM) [1, 2, 3], the Theory of Planned Behavior (TPB) [4] and the Diffusion of Innovation (DOI) theory [5]. However, many authors (e.g. (Pedersen, Per E. and Nysveen, 2002; Pedersen, Per E. and Ling, 2003; Yu, Liu et al., 2003; Kim, Chan et al., 2005; Nysveen, Pedersen et al., 2005) have pointed out that traditional adoption models are insufficient to gain a comprehensive explanation of the factors that affect individuals' intentions to adopt or reject the use of mobile commerce services [6, 7, 8, 9, 10].

In this paper a new approach for assessing potential adoption of future mobile services is proposed based on the Technology Acceptance Model (TAM). The TAM model has already been applied for analyzing the adoption and acceptance of existing mobile services. Several studies have shown that TAM can explain the influence of different factors on the customers' intention to use mobile services [9, 10, 11, 12, 13]. However, one major problem in the telecom industry is to assess the potential future adoption of mobile services that are not launched on the market yet.

The goal of the research presented in this paper is the exploration of the applicability of the TAM model for evaluation of potential future mobile services. The potential intention to use the mobile services described by the scenarios was evaluated based on an adjusted TAM model for mobile services. The analysis revealed that TAM is a suitable approach to evaluate the potential intention to use mobile services. The results also showed that the most influential factor is the perceived value of the future services.

The paper is structured as follows. In the beginning of section 2 collect and integrate the reference for related theories. In section 3 the research model and hypotheses is described followed by the results in section 4. In section 5 and 6, the implications and the demand for future research are discussed.

II. LITERATURE REVIEW

[1] Corresponding author: Tel.: (86)138-1872-1157; E-mail: caohao309@163.om

doi:10.4304/jcp.5.5.799-806

A. Theory of Planned Behavior

TPB underlying the effort of TRA has been proven successful in predicting and explaining human behavior across various information technologies [4]. According to TPB, a person's actual behavior in performing certain action is directly influenced by his or her behavioral intention and in turn, jointly determined by attitude, subjective norm and perceived behavioral control toward performing the behavior. Behavioral intention is a measure of the strength of one's willingness to try and exert while performing certain behavior. Attitude (A) explains the feeling of a person's favorable or unfavorable assessment regarding the behavior in question. Furthermore, a favorable or unfavorable attitude is a direct influence to the strength of behavior al beliefs about the likely salient consequences. Accordingly, attitude (A) is equated with attitudinal belief (Bi) linking the behavior to a certain outcome weighted by an evaluation of the desirability of that outcome (E_i) in question, i.e.A = $\sum B_i E_i$. Subjective norm (SN) expresses the perceived organizational or social pressure of a person while intending to perform the behavior in question. In other word, subjective norm is relative to normative beliefs about the expectations of other persons. It can be depicted as individual's normative belief (NB_i) concerning a particular referent weighted by motivation to comply with that referent (MC_i) in question, i.e. SN $= \sum NB_i MC_i$.

Perceived behavioral control (PBC) reflects a person's perception of ease or difficulty toward implementing the behavior in interest. It concerns the beliefs about presence of control factors that may facilitate or hinder to perform the behavior. Thus, control beliefs about resources and opportunities are the underlying determinant of perceived behavioral control and it can be depicted as control beliefs (CBi) weighted by perceived power of the control factor (PFi) in question, i.e. PBC= $\sum CB_i PF_i$. In sum, grounded on the effort of TRA, TPB is proposed to eliminate the limitations of the original model in dealing with the behavior over which people have incomplete volitional control [4]. In essence, TPB differs from TRA in its addition of the component of perceived behavior control.

B. Technology Acceptance Model

Davis used TRA as its theoretical foundation to discuss the correlation among cognitive, emotion and application of technology, and then developed a framework as Technology Acceptance Model (TAM), shown as Fig. 1.

This model provides a theoretical foundation to understand how external variables influence the inner beliefs, attitude, and intention of users, and then affect the use of technology. The purpose of TAM is to provide an explanation toward the acceptance of technology which explains users' behavior on accepting new information technology, and analyzes the factors that influence their attitude toward using new information technology [1, 2, 3].

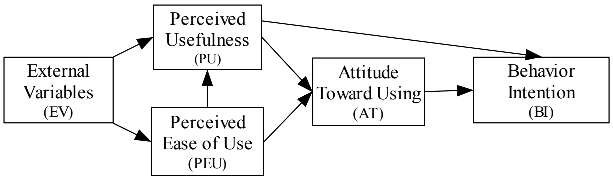

Perceived Usefulness (PU)
External Variables (EV)
Attitude Toward Using (AT)
Behavior Intention (BI)
Perceived Ease of Use (PEU)

Figure 1. Technology Acceptance Model

TAM further suggests that two beliefs–perceived usefulness and perceived ease of use–are instrumental in explaining the variance in the intention of users. Perceived usefulness is defined as the extent to which a person believes that using a particular system will enhance his or her job performance, and perceived ease of use is defined as the extent to which a person believes that using a particular system will be free of effort. Among the beliefs, perceived ease of use is hypothesized to be a predictor of perceived usefulness. Information system researchers have investigated and replicated the TAM, and agreed that it is valid in predicting an individual's acceptance of various corporate IT [14, 15]. However, the TAM's fundamental constructs do not fully reflect the specific influences of technological and usage-context factors that may alter user acceptance [2]. Thus, prior studies have extended the TAM with constructs such as perceived playfulness [16, 17], compatibility [18], perceived user resources [19], trust [20], perceived credibility [21, 22] and trustwor-thiness [23]. Venkatesh et al. reviewed eight popular models and combined them to the Unified Theory of Acceptance and Use of Technology (UTAUT) to explain the acceptance of information systems [24]. Most of the research in technology acceptance was done on technologies that were introduced into organizations and could therefore only partially describe the completely voluntary usage of technologies such as the mobile phone by independent end users. Only recently the technology acceptance theory was applied to Mobile Services.

C. Extension of TAM model

In order to apply the TAM model, it is necessary to operational its components in accordance to the specific characteristics of the technology under consideration. The original TAM suggested by Davis focused on the two components perceived usefulness and perceived ease of use. Even though the findings of Davis were considered, it was not possible to apply them to the evaluation of future broadcasting services, because they are oriented towards adoption of technologies in companies. The mobile services considered in this paper are entertainment services. Thus, the "perceived usefulness" and "perceived easy of use" had to be operational in a suitable way in order to encounter the specific entertainment characteristics of mobile broadcasting services.

Various studies have found that trust is strongly associated with attitude towards products and services and towards purchasing behaviors. In accordance with Kaasinen another specific and important factor influencing the acceptance of mobile services is trust [25, 26]. Also, Keat & Mohan suggested adding a component describing the trust to the TAM. Trust is a combination of level of familiarity, the company reputation, factual signals, and the quality of experience [27]. Kaasinen furthermore combined the specific components of TAM for mobile services in a new version of TAM dedicated to mobile services. Kaasinen modified the value component (from perceived usefulness) and added the components trust and perceived ease of adoption [26].

Many different definitions of trust have been proposed [28,29],We defined trust as party trust and control trust, based on the study of Tan & Theon, Jieun Yu et al. [30,31]. According to their study, transaction trust consisted of trust in the other party and trust in the control mechanisms. Party trust was subjective trust in a transaction with another party and had both an action and information perspective. We limited party trust to a belief in a mobile services provider's after-services. Control trust was then trust in control mechanisms that ensured the successful performance of the transaction. We limited this to stable payment, system security, and personal information protection. Since interactive information exchange takes place in the use of mobile services, perceived variables about system security or personal information protection seriously affect adoption [20]. Gefen suggested that trust in an e-commerce vendor increased a user's intention to use the vendor's web site and was the most efficient factor for reducing uncertainty [20]. Doney and Cannon showed that customer trust was related to intention to use the vendor in the future [33]. Grazioli and Jarvenpaa found that customers' attitude was determined by trust in the context of an Internet shopping mall[34]. Therefore, trust would be an important factor in influencing consumers' attitudes and BI toward the adoption of mobile services.

With the increasingly high penetration rate of Internet applications, people are anxious about the diverse types of risks presented when engaging in online activities or transactions. When customers are uncertain about product quality, brands and online services they may worry about an unjustifiable delay in product delivery, providing payment without receiving the product and other illegal activities and fraud. The theory of perceived risk has been applied to explain consumer's behavior in decision making since the 1960s [35]. The definition of perceived risk has changed since online transactions became popular. In the past, perceived risks were primarily regarded as fraud and product quality. Today, perceived risk refers to certain types of financial, product performance, social, psychological, physical, or time risks when consumers make transactions online [35,36].Other research also indicated that perceived risk is an important determinant of consumers' attitude toward online transactions [35,37]. Since intention to use a mobile phone for transactions involves a certain degree of uncertainty, perceived risk is incorporated as a direct antecedent of behavioral intention to use.

According to behavioral decision theory, the perceived cost pattern is significant to both perceived usefulness and ease of use. As Chen and Hitt and Plouffe et al. pointed out, consumers must deal with non-negligible costs in switching between different brands of products or relative services invarious markets. Transitioning from wired electronic services (ES) to mobile services (MS) implies some additional expenses. Equipment costs, access cost, and transaction fees are three important components that make MS use more expensive than wired ES. Furthermore, frustrating experiences, such as slow connections, poor quality, out-of-date content, missing links, and errors have infuriated online users [34, 35].

Unfortunately, consumers must pay for all these frustrations. Some researchers suggested that MS providers should find solutions that reduce the costs and entice present and new customers to access portals anytime, from anywhere [35]. Undoubtedly, the anticipation is that these early investments will lead to a long-term stream of profits from loyal customers, and that this will make up for the expense. Otherwise, MS will not thrive because users can obtain the same information or results through alternative solutions.

Previous research has suggested that trust-related constructs, cost-constructs and risk-related constructs should be the critical antecedents of the behavioral intention to use IS. Integrating these perspectives and empirically examining the factors that build usage intention in an m-service context that lacks typical human interaction, can advance our understanding of these constructs and their link to m-service adoption behavior. Wang et al. define trust and risk is Perceived Credibility, perceived credibility is defined as the extent to which a person believes that using m-service will be free of security and privacy threats [21, 22]. Perceived credibility was also found to have a significant positive influence on the behavioral intentions to use online banking [21], electronic tax filing [21], electronic learning [38] and m-banking [39]. In general, the perceived credibility that people have in the ability of the m-service system to conclude their transactions securely and to maintain the privacy of their personal information, affects their voluntary acceptance of m-service. Based on the previous theory and then developed a framework as an extension of TAM model, shown as Fig. 2.

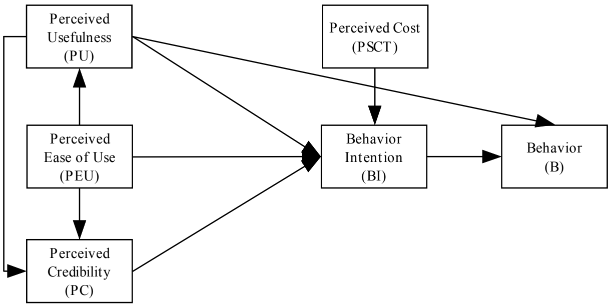

Perceived Usefulness (PU)
Perceived Cost (PSCT)
Perceived Ease of Use (PEU)
Behavior Intention (BI)
Behavior (B)
Perceived Credibility (PC)

Figure 2. Extension of TAM Model

III. Research Model and Hypotheses

The research model is shown in Fig. 3. In the extended model, like many other studies of TAM, our model integrates the key variables of TAM (perceived usefulness and perceived ease of use), TPB (perceived behavioral control), and "attitudes" construct has been removed for simplification. The proposed constructs and hypotheses are supported by prior studies in information systems literature.

Based on this, the following hypotheses are proposed:

H1. Perceived usefulness will have a positive effect on the behavioral intentions.

H2a. Perceived ease of use will have a positive effect on the behavioral intentions.

H2b. Perceived ease of use will have a positive effect on the perceived usefulness.

H3a. Perceived behavioral control will have a positive effect on behavioral intention.

H3b. Perceived behavioral control will have a positive effect on perceived ease of use.

H4a. Perceived ease of use will have a positive effect on perceived credibility.

H4b. Perceived credibility will have a positive effect on perceived usefulness.

H4c. Perceived credibility will have a positive effect on behavioral intention.

H5: Perceived cost will have a positive effect on behavioral intention.

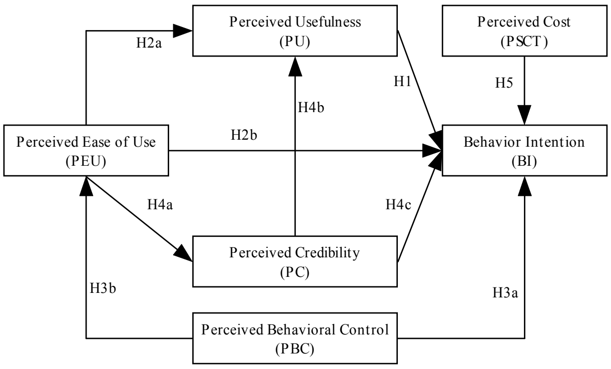

Perceived Usefulness (PU)
Perceived Cost (PSCT)
H2a
H1
H5
H4b
Perceived Ease of Use (PEU)
H2b
Behavior Intention (BI)
H4a
H4c
Perceived Credibility (PC)
H3b
H3a
Perceived Behavioral Control (PBC)

Figure 3. Research Model

IV. Test and Result

A. Data Collection

We conducted an online survey to verify our research model. After data-filtering to eliminate invalid responses, we had received 228 questionnaires. Most questions in our questionnaire were taken from prior studies that had proved their validity and reliability. Each item of the questionnaire was assessed using a 5-point Likert scale with end points of 'strongly disagree' and 'strongly agree'. Accordingly, the first part is basic information. This part of questionnaire was used to collect basic information about respondents' characteristics including gender, age, education, occupation, and experience using online banking. The second part of questionnaire was developed based on the constructs of perceived usefulness, perceived ease of use, perceived credibility, perceived behavior control, perceived cost, and behavior intention to use. Before conducting the main survey, we performed a pre-test to validate the instrument. The pre-test involved 10 respondents who have more than 1 years experience using mobile service. Respondents were asked to comment on the length of the instrument, the format, and the wording of the scales. Therefore, the instrument has confirmed content validity.

B. Reliability Analysis

Our data analysis was conducted using SPSS 13.0. Internal consistency reflects the stability of individual measurement items across replications from the same source of information; it was assessed by computing Cronbach's α whose coefficients for the eight constructs were above 0.7, indicating a reasonable level of internal consistency among the items making it up, Show as Table I.

TABLE I
CONSTRUCTS MEAN, STANDARD DEVIATION AND INTERNAL CONSISTENCY RELIABILITY

Constructs(Items)	Mean	Standard deviation	Cronbach's α
Perceived usefulness			0.76
PU1:Using mobile services would improve my performance in conducting services transactions	4.20	0.80	
PU2:Using mobile services would make it easier for me to conduct services transactions	4.34	0.78	
PU3:I would find mobile services useful in conducting my service transaction	4.41	0.78	
Perceived ease of use			0.78
PEU1:Learning to use mobile services is easy for me	4.03	1.00	
PEU2:It would be easy for me to become skillful at using mobile services	3.80	1.02	
PEU3:I would find mobile services easy for me	4.00	0.87	
Perceived credibility			0.72
PC1:Using mobile services would not divulge my personal information	3.32	1.05	
PC2:I would find mobile services secure in conducting my services transactions	3.52	1.02	
Perceived behavioral control			0.71
PBC1:In my experience and expectations, I believe to use mobile services will not have much difficulty	3.70	0.93	
PBC2:I think I have more resources and opportunities, expected fewer obstacles, to act to control behavior on the perception is stronger	3.90	0.85	
Perceived cost			0.81
PCST1:It would cost a lot to use mobile services	3.30	0.94	
PCST2:There are financial barriers to my using mobile services(e.g., having a pay for handset and communication time)	3.40	0.90	
Behavioral intention			0.88
BI1:Assuming that I have access to mobile services systems, I intend to use them	3.65	0.94	
BI2:I intend to increase my use of mobile services in the future	3.50	0.99	

C. Validity Analysis

A confirmatory factor analysis using LISREL 8.3 was conducted to test the measurement model. Nine common model-fit measures were used to assess the model's overall goodness of fit: the ratio of χ^2 to degrees of freedom (d.f.), normalized fit index (NFI), non-normalized fit index (NNFI), comparative fit index (CFI), goodness-of-fit index (GFI), adjusted goodness-of-fit index (AGFI), and root mean square error of approximation (RMSEA). As shown in Table II, all the model-fit indices exceeded their respective common acceptance levels suggested by previous research, thus demonstrating that the measurement model exhibited a fairly good fit with the data collected. Therefore, we could proceed to evaluate the psychometric properties of the measurement model in terms of reliability, convergent validity and discriminant validity [27].

TABLE II
THE MODEL-FIT INDICES

Fit Indices	Recommended value	Measurement model	Structural model
χ^2/d.f.	≤3.00	1.494	1.805
NFI	≥0.90	0.955	0.931
NNFI	≥0.90	0.916	0.909
CFI	≥0.90	0.942	0.925
GFI	≥0.90	0.973	0.954
AGFI	≥0.80	0.980	0.963
RMSEA	≤0.08	0.044	0.056

Composite reliability for all the factors in our measurement model was above 0.70. The average extracted variances were all above the recommended 0.50 level [10], which meant that more than one-half of the variances observed in the items, were accounted for by their hypothesized factors. Convergent validity can also be evaluated by examining the factor loadings; composite reliability and average variance extracted from the confirmatory factor analysis (see Table III). Following Fornell,C. & Larcker, D.F.(1981) recommendation: (1)factor loadings greater than 0.50 were considered very significant. All of the factor loadings of the items in the research model were greater than 0.70. (2) Composite reliability greater than 0.8 were considered very significant. All of the Composite reliability of the items in the research model were greater than 0.8. (3) Average variance extracted greater than 0.5 were considered very significant [39]. All of the Average variance extracted of the items in the research model were greater than 0.7. Accordingly, all factors in the measurement model had adequate reliability and convergent validity.

Reliability and convergent validity of the factors were estimated by composite reliability and average variance extracted (see Table IV). The composite reliabilities can be calculated as follows: (square of the summation of the factor loadings)/{(square of the summation of the factor loadings)+(summation of error variables)}, where the factor loadings are obtained directly from the program output, and the error variables is the measurement error for each indicator. The interpretation of the composite reliability is similar to that of Cronbach's alpha, expect that it also takes into account the actual factor loadings, rather than assuming that each item is equally weighted in the composite load determination.

TABLE III
FACTOR LOADINGS, COMPOSITE RELIABILITY AND AVERAGE VARIANCE EXTRACTED

Variables	Constructs	Factor loadings	Composite reliability	Average variance extracted
PU	PU1	0.86	0.90	0.83
	PU2	0.82		
	PU3	0.91		
PEU	PEU1	0.83	0.83	0.80
	PEU2	0.78		
	PEU3	0.80		
PC	PC1	0.92	0.88	0.83
	PC2	0.95		
PBC	PBC1	0.80	0.85	0.78
	PBC2	0.89		
PCST	PCST1	0.82	0.81	0.72
	PCST2	0.98		
BI	BI1	0.92	0.87	0.84
	BI2	0.95		

TABLE IV
COMPOSITE RELIABILITY AND AVERAGE VARIANCE EXTRACTED

Variables	1	2	3	4	5	6
PU	1.00					
PEU	0.36**	1.00				
PC	0.51**	0.42**	1.00			
PBC	0.16	0.40**	0.22**	1.00		
PCST	0.40**	0.33**	0.26**	0.45**	1.00	
BI	0.02	0.01	0.09	0.23**	0.13	1.00

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

KMO Measure of Sampling Adequacy: 0.818

Bartlett's Test of Sphericity: 405.144; Significance: .000

D. Structural Model

A similar set of indices was used to examine the structural model. Comparison of all fit indices with their corresponding recommended values provided evidence of a good model fit. Thus, we proceeded to examine the path coefficients of the structural model. Properties of the causal paths, including standardized path coefficients, t-values and variance explained, for each equation in the hypothesized model, are presented in Fig. 4.

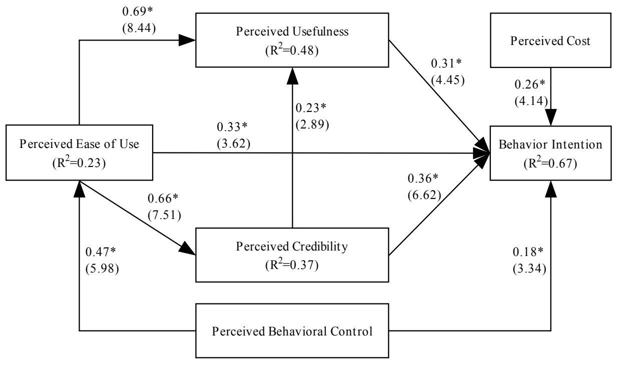


Figure 4. Hypotheses testing results

* $P < 0.05$; t-values in parentheses.

As expected, hypotheses H1, H2b, H4c, H3a and H5 were supported in that perceived usefulness, perceived ease of use, perceived credibility, perceived behavioral control and perceived cost all had a significant effect on behavioral intention. Altogether, they accounted for 67% of the variance in behavioral intention with perceived usefulness (β=0.31) contributing to intention nearly perceived ease of use (β=0.33), perceived credibility (β=0.36), perceived behavioral control (γ=0.18) and perceived cost (γ=0.26). In addition, hypotheses H2, H3, H6 and H7 were also supported. Perceived behavioral control was found to have a significant influence on perceived ease of use (γ=0.47). Both perceived ease of use and perceived credibility had a significant effect on perceived usefulness (γ=0.69 and β=0.23, respectively). Thus, perceived ease of use had a significant effect on perceived credibility (γ=0.66).

V. Conclusions and Discussions

Prior studies have found that TAM appears to be superior to TPB in explaining behavioral intention of using an IS, and that the decomposed TPB model, which integrates TPB and TAM, is better than TAM but the difference is not substantial [17].

This study based on the TAM and TPB presents an extended TAM that integrates trust related construct (‘perceived credibility’) and two resource-related constructs (‘perceived self-efficacy’ and ‘perceived cost’) into the TAM to analyzing adoption behavior of mobile commerce. The proposed model was empirically tested using data collected from a survey of mobile commerce consumers. The structural equation modeling technique was used to evaluate the causal model and confirmatory factor analysis was performed to examine the reliability and validity of the measurement model. Our findings indicated that all variables except perceived ease of use significantly affected users’ behavioral intention. Compared with prior studies integrating TAM and TPB, the findings of this study strongly suggest that our model with only five independent constructs has a higher ability to predict and explain the behavioural intention of users to use an IS. The variance in intention explained (R^2) in our study was 67%. The results of this paper are to provide necessary reference to promote mobile commerce in practice. The results show that there are some significant factors affect consumers’ using mobile commerce as follows. According to the results of this research, mobile commerce companies can affect consumers' willingness of using mobile commerce, to increase the number of mobile commerce users.

1) Perceived Usefulness: Mobile commerce companies can strengthen the system function; improve consumer perception of mobile commerce usefulness, to enhance behavior intention of using mobile commerce. 2) Perceived Ease of Use: Mobile commerce companies can strengthen the system ease of the use, and enhance behavior intention of using mobile commerce by increasing the system ease of the use. 3) Perceived Credibility: Mobile commerce companies can strengthen the transaction security and privacy protection abilities in the mobile commerce, to reduce the risk of consumer perception. 4)Perceived Behavioral Control: Mobile commerce companies should keep abreast of consumers' results generated by using mobile commerce and consumers' response to the results. 5) Perceived Cost: Mobile commerce companies should make the mobile commerce cost as low as possible.

Although this paper makes certain achievements, it also faces many restrictions. First, due to the time limit, we only collected a small quantity of questionnaire, so there are some shortcomings in the sample amount; more samples should be tested in the follow-up research. Second, this paper only researches the mobile commerce, it should be prudent and careful to apply the research results to the whole mobile business, and the research results can’t be explained excessively. Subsequent research can investigate the application in other mobile service to enhance the development of the mobile business theory. In the follow-up research study, we can join other important factors to the three theory models for further research to improve the models.

VI. Limitations and Future Research

In terms of academic, TAM is used to realize users’ behaviors to accept new IT, and try to analyze factors that influence users to accept new IT; it is known from the research that the TAM’s path relation among factors is workable under mobile service, and has certain explanatory capability, which shows consumer’s highly attitude and intention to accept mobile service is conform to the path relation of original TAM.

In terms of industry, the research results enable mobile service providers to realize consumers’ conceptual psychology factor toward mobile service, which show that even though consumers have highly attitude and intention toward mobile service, the actual use is not as expected; therefore, services providers could make further research for the causes to find out the real causes why consumers do not frequently used and then improve use condition.

Due to time and source limitations, the research cannot be done as a long-term research, and did not have the chance to understand the individual purpose for using mobile service. The research is done regionally rather than nationally due to various restrictions. Besides, in order to let the research proceed smoothly, the selection for services did not cover all types of mobile commerce services, but to list a few most commonly used services for survey.

On the other hand, due to part of the measurements were sent to the interviewees, their willingness for replying the measurement may not be high, and if the one who replied the measurement was not the one we would like to ask, it would influence the result as well. Furthermore, the interviewees who took their time replying the measurements indicate that they are interested in the topic we enquired, in other words, the recovered measurements might be done a group of people who have interest in this field which might create bias for the result. Moreover, the questions were simplified in order to reduce the pages of the measurement and make the interviewees feel easier to reply which might influence the result as well.

From the experiences we had this time, we suggested that other researchers may discuss the same topic from the following directions. Due to the fact that the samples collected in this research were limited, the result cannot be generalization. For future research, the samples could be selected from national-wide, or every industry and long-term follow-up study could be made to understand deeply about the consumers' behaviors. TAM was adopted for this research as to understand the consumers' attitude and behavioral intention toward mobile service, in the future, other behavioral theories could be adopted or introduce for proceeding future research as well. Moreover, other variables that would influence the actual use of mobile service, stated by Liao and Cheung (2001), could include the price, using experiences, brand, education, sense for using information technology, and communication quality for wireless network etc., all of the above items could let us understand more about the variables that influence consumers' actual use toward mobile commerce [40].

Acknowledgment

The authors would like to thank the guest editors and two anonymous reviewers for their valuable comments and insightful suggestions. This research was substantially supported by the Science and Technology Commission of Shanghai Municipality "Research and Development of Public Service Platform Construction Project" under grant number 06DZ22924, Shanghai Science and Technology Development Fund Soft Science Doctoral Thesis Research under grant number 200906017 and Suzhou philosophy and social science research topics under grant number 09-C-45.

References

[1] F. D. Davis, A "Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results," Doctoral Dissertation, MIT Sloan School of Management, Cambridge, MA, 1986.

[2] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "User Acceptance of Computer Technology: A Comparison of two Theoretical models," Management Science, 1989, vol. 35, pp. 982-1003.

[3] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," MIS Quarterly, 1989, vol. 13, pp. 319-340.

[4] I. Ajzen, "From intentions to actions: A theory of planned behavior. In Kuhl J.and Beckmann J.(eds.). Action Control: From Cognition to Behavior," New York: Springer-Verlag. 1985, vol. 3, pp. 11-39.

[5] E. M. Rogers, "Diffusion of innovations" / Everett M. Rogers. New York., NY:, Free Press, 2003.

[6] P. E. Pedersen, H. Nysveen, "Using the theory of planned behavior to explain teenagers' adoption of text messaging services," Working Paper, Agder University College, 2002.

[7] P. E. Pedersen, R. Ling, "Modifying adoption research for mobile Internet service adoption: Cross-disciplinary interactions," The 36th Hawaii International Conference on System Sciences (HICSS'03), Hawaii, IEEE Computer Society, 2003.

[8] J. Lu, C. S. Yu, C. Liu, J. E. Yao, "Technology acceptance model for wireless Internet," Internet Research, 2003, vol. 13, pp. 206-222.

[9] H. Nysveen, P. E. Pedersen, and H. Thorbjornsen, "Explaining intention to use mobile chat services: moderating effects of gender." Journal of Consumer Marketing, 2005, vol. 22, pp. 247-256.

[10] H. Nysveen, P. E. Pedersen, and H. Thorbjornsen, "Intentions to Use Mobile Services: Antecedents and Cross-Service Comparisons," Academy of Marketing Science. Journal, 2005, vol. 33, pp. 330-346.

[11] J. H. Cheong, M. C. Park, "Mobile internet acceptance in Korea," Internet Research, 2005, vol. 15, pp. 125-140.

[12] S. Y. Hung, C. Y. Ku, and C. M. Chang, "Critical factors of WAP services adoption: an empirical study," Electronic Commerce Research and Applications, 2003, vol. 2, pp. 42-60.

[13] J. Lu, C. S. Yu, C. Liu, and J. E. Yao, "Technology acceptance model for wireless Internet," Internet Research, 2003, vol. 13, pp. 206 - 222.

[14] D. A. Adams, R. R. Nelson, and P. A. Todd, "Perceived usefulness, ease of use, and usage of information technology: a replication," MIS Quarterly, 1992, vol. 16, pp. 227–247.

[15] A. H. Segars, V. Grover, "Re-examining perceived ease of use and usefulness: a confirmatory factor analysis," MIS Quarterly, 1993, vol. 17, pp. 517–525.

[16] J. W. Moon, Y. G. Kim, "Extending the TAM for a World-Wide-Web context," Information & Management, 2001, vol. 38, pp. 217-230.

[17] P. Y. K. Chau, P. J. H. Hu, "Investigating healthcare professionals' decisions to accept telemedicine technology: an empirical test of competing theories," Information & management, 2002, vol. 39, pp. 297-311.

[18] K. Mathieson, E. Peacock, and W. W. Chin, "Extending the technology acceptance model: the influence of perceived user resources," DATA BASE for Advances in Information Systems, 2001, vol. 32, pp. 86–112.

[19] D. Gefen, E. Karahanna, and D. W. Straub, "Trust and TAM in online Shopping: An Integrated Model," MIS Quarterly, 2003, vol. 27, pp. 51-90.

[20] Y. S. Wang, Y. M. Wang, H. H. Lin, and T. I. Tang, "Determinants of user acceptance of internet banking: An empirical study," International Journal of Service Industry Management, 2003, vol. 14, pp. 501–519.

[21] Y. S. Wang, H. H. Lin, and P. Luarn, "Predicting consumer intention to use mobile service," Information Systems Journal, 2006, vol. 16, pp. 157-179.

[22] L. Carter, F. Bélanger, "The utilization of e-government services: citizen trust, innovation and acceptance factors," Information Systems Journal, 2005, vol. 15, pp. 5–26.

[23] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: toward a unified view," MIS Quarterly, 2003, vol. 27, pp. 425–478.

[24] E. Kaasinen, V. Ikonen, A. Ahonen, V. Anttila, M. Kulju, J. Luoma, and R. Södergård, "Products and Services for Personal Navigation . Usability Design. Part III. Case Studies and Usability Guidelines," Publications of the NAVI programme. www.vtt.fi/virtual/navi [online, cited 4.1.2005], 2003.

[25] E. Kaasinen, V. Ikonen, A. Ahonen, V. Anttila, M. Kulju, J. Luoma, and R. Södergård, "Products and Services for Personal Navigation . Classification from the User's Point of View," Publications of the NAVI programme. www.vtt.fi/virtual/navi [online, cited 4.1.2005], 2002.

[26] T. K. Keat and A. Mohan, "Integration of TAM based electronic commerce models for trust," J. Am. Acad. Bus., 2004, vol. 5, pp. 404–410,.

[27] D. M. Koo, "An investigation on consumer's internet shopping behavior explained by the technology acceptance model," Journal of Korean Management Information System, 2003, vol. 14, pp. 141–170

[28] K. N. Chang, W. J. Yang, Y. J. Park, "An analysis on trust factors of B2C electronic commerce," Informatization Policy, 2002, vol. 9, pp. 3–17.

[29] Y. H. Tan, W. Theon, "Toward a generic model of trust for electronic commerce," International Journal of Electronic Commerce, 2000-2001, vol. 5(2), pp. 61–74.

[30] J. Yu, I. Ha, M. Choi, and J. Rho, "Extending the TAM for a t-commerce," Information & Management, 2005, vol. 42, pp. 965–976.

[31] S. Grazioli, S. L. Jarvenpaa, "Perils of Internet fraud: an empirical investigation of deception and trust with experienced Internet," IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans, 2000, vol. 30, pp. 395–410.

[32] P. M. Doney, J. P. Cannon, "An examination of the nature of trust in buyer–seller relationships," Journal of Marketing, 1997, vol. 61, pp. 35–51.

[33] E. Constantinides, "The 4S Web-marketing mix model, electronic commerce research and applications," Elsevier Science, 2002, vol. 1, pp. 57–76.

[34] J. H. Wu, S. C. Wang, "What drives mobile commerce? An empirical evaluation of the revised technology acceptance model," Information & Management, 2005, vol. 42, pp. 719–729.

[35] S. M. Forsythe, B. Shi, "Consumer patronage and risk perceptions in Internet shopping," Journal of Business Research, 2003, vol. 56, pp. 867–875.

[36] J. Cho, "Likelihood to abort an online transaction: influences from cognitive evaluations, attitudes, and behavioral variables," Information & Management, 2004, vol. 41, pp. 827–838.

[37] C. S. Ong, J. Y. Lai, Y. S Wang, "Factors affecting engineers' acceptance of asynchronous e-learning systems in high-tech companies," Information and Management, 2004, vol. 41, pp. 795-804.

[38] P. Luarn, H. H. Lin, "Toward an understanding of the behavioral intention to use mobile banking,"Computers in Human Behavior, 2005, vol. 21, pp. 873-891.

[39] C. Fornell, D. F. Larcker, "Structural Equation Model with Unobservable Variables and Measurement Error: Algebra and statistics," Journal of Marketing Research, 1981, vol. 18, pp. 382-389.

[40] Z. Liao, M. T. Cheung, "Internet-based e-shopping and consumer attitudes: an empirical study", Information & Management, 2001, Vol.38, pp. 299-306.

Sun Quan has a Master Degree in Financial Engineering from The Xi'an Jiao Tong University. He is a lecturer in the Department of Business at the Suzhou Vocational University, China. His current research relates to financial engineering, information systems, business strategy and mobile commerce.

Cao Hao is a doctoral candidate in the School of Economics and Management, Chinese Academy of S& T Management, at Tongji University, China. His current research relates to management of technology, science and technology finance, and financial engineering.

You Jianxin is a professor in the School of Economics and Management, Chinese Academy of S& T Management, at Tongji University, China. His current research relates to quality control, business administration and management of technology.

Provable Secure Generalized Signcryption

Xu an Wang[1], Xiaoyuan Yang[1] and Jindan Zhang[2]
[1] Key Laboratory of Information and Network Security,
Engneering College of Chinese Armed Police Force, 710086, P. R. China
[2] Department of Electronic Information,
Xianyang Vocational Technical College, 712000, P. R. China
E-mail:wangxahq@yahoo.com.cn

Abstract— **Generalized signcryption which proposed by Han is a new cryptographic primitive which can work as an encryption scheme, a signature scheme or a signcryption scheme [5]. However,the security proof in their paper is uncorrect.our contribution are as following:First we give security notions for this new primitive.Secnond,we give an attack to [4]which is the first vision of [5] and propose an improved generalized signcryption scheme. Third, we give correct proofs for this new scheme.**

Index Terms – **generalized signcryption, provable security, attack, security notions.**

I. Introduction

Along with developments of information society, security requirements for applications are usually both confidentiality and authentication. And these requirements have given birth of new research fields in cryptography, that is, how to combine confidentiality and authentication properly. A lot of work has been done in this field, such as how to encrypt message by block cipher properly to achieve authentication or how to combine ciphertext with signature properly to achieve authentication [1], [8]. Totally we can divide the work into three types: Encryption then Sign, Sign then Encryption, Encryption and Sign. In 1997, Zheng proposed a new cryptographic primitive: Signcryption [2]. The idea is compressing two independent operations (encryption and signature) in one operation (signcryption). There are three advantages from this transformation: reducing the steps needed by encryption and signature(less computation complexity); reducing length of ciphertext produced by encryption and signature(less communication complexity); reducing two modules of encryption and signature to one module of signcryption(less implementation complexity). Since then, a lot of research results have come out. We can see SCS-DSA, SCS-KCDSA signcryption scheme based on Discrete Logarithm problem, RSA-TBOS signcryption scheme based on Integer Factoring [6], ECSCS signcryption scheme based on elliptic curve [7], identity based signcryption scheme based on pairings.

In 2006, Han proposed a new primitive generalized signcryption [3]. The idea of this new primitive is still reducing, but this time, what's reducing is not the computation complexity or communication complexity, but the implementation complexity. Imagine this scenario, two users want to communicate safely. Sometimes they need both confidentiality and authentication, sometimes they just need confidentiality, and sometimes they just need authentication. If we adopt signcryption in this scenario, we must preserve module of encryption and module of signature for solely needing confidentiality or authentication. If we do not care very much about speed, we gain no remarkable advantage for adopting signcryption. Furthermore, adding something new to an established system seems no easy. But if we can embed encryption and signature in the signcryption module, we can easily encrypt or sign or signcrypt by only one module.

A. Motivation

Generalized Signcryption is the one which fits this goal. Generalized Signcryption is a new primitive which can work as an encryption scheme, a signature scheme, or a signcryption scheme. Maybe this can broaden the application range of signcryption.We must point out here that Generalized Signcryption can not substitute of encryption or signature. But it fit some particular application perfectly.

B. Related Works

Actually, the generalized signcryption concept is not new, it has been mentioned in Zheng's original paper [2]. In [20] Boyen et al proposed a mulitpurpose signcryption which they called as a swiss armed knife,the motivation is similar to our's. In [10], [11],Dodis et al proposed a versatile padding schemes which can perfectly played as an encryption or signature or signcryption scheme,The technique in their paper is padding message before processing. In the two extremities, the scheme turns to be OAEP-padding and PSS-padding. In the non-extremity, the scheme turns to be signcryption,furthermore, they prove their result is optimal, but they do not propose the generalized signcryption concept which is [5] main contribution.

C. Our contribution

However, [5] do not give the formal model for this new primitive and unfortunately the security proof for their scheme is uncorrect.Actually, all the papers [10], [11], [20]mentioned above do not consider formal security model for this multi-functionality cryptographic primitive. In this paper, we reconsider this new primitve thoroughly. our contribution are as following:First we give security notions for this new primitive.Secnond,we give an attack to [4]which is the first vision of [5] and propose an improved generalized signcryption scheme. Third, we give correct proofs for this new scheme.

doi:10.4304/jcp.5.5.807-814

D. Organization

The paper is organized as following: In the second section, we give new formal model for this new primitive which is based on the theory of provable security [14]–[19]. In the third section, we give an attack to the origin scheme in [4],which is the first vision of [5],and we give an improved scheme by give little change to the original scheme.In the forth section,we give formal correct proofs for this improved Generalized Signcryption scheme, which implies scheme in [5] be secure.We give our conclusion in the last section.

II. Generalized Signcryption and Its Security Notions

A. Definition of Generalized Signcryption and a Concrete Scheme ECGSC

Generalized Signcryption is a signcryption with more flexibility and practicability. It provides double Functions when confidentiality and authenticity are required simultaneously, and provides single Encryption or signature function when confidentiality Or authenticity is required only without any amended and additional computation. Namely, a generalized signcryption scheme will be equivalent to a signature scheme or an encryption scheme in special cases. Hence, a generalized signcryption will work in modes: signcryption, signature-only, and encryption-only.

Definition 1: Given a normal secure signature scheme $SIG = (Gen, Sig, Ver)$ where Gen is a key generation algorithm, $\tau \leftarrow Sig(m, SDK_S)$,$(T, \perp) \leftarrow Ver(\tau, VEK_S)$, a normal secure encryption scheme $ENC = (Gen, Enc, Dec)$ where Gen is the same algorithm as SIG's Gen,$\varepsilon \leftarrow Enc(m, VEK_R)$,$m \cup \{\perp\} \leftarrow Dec(\varepsilon, SDK_R)$ and a normal secure signcryption scheme $SC = (Gen, Sc, Usc)$ where Gen is the same algorithm as SIG's Gen,$w \leftarrow Sc(m, SDK_S, VEK_R)$,$(m \cup \{\perp\}) \cup (T, \perp) \leftarrow Usc(w, SDK_R, VEK_S)$.A generalized signcryption scheme $GSC = (Gen, Gsc, Ugsc)$ should be constructed satisfying the following:

1) KeyGen: Must be the same algorithm as Gen.
2) Generalized Signcryption: For $m \in M$,$w \leftarrow Gsc(m, SDK_S, VEK_R)$. When S is a special value, Gsc(m, SDK_S,VEK_R)=Enc(m, VEK_R); When R is a special value,Gsc(m,SDK_S,VEK_R)=Sig(m,SDK_S); When S and R are both not special values, Gsc(m, SDK_S, VEK_R)=Sc(m, SDK_S,VEK_R);
3) Generalized Unsigncryption: For $w \in \mathcal{C}$,$(m \cup \{\perp\}) \cup (T, \perp) \leftarrow Ugsc(w, SDK_R, VEK_S)$. When S is a special value,$Ugsc(w, SDK_R, VEK_S) = Dec(\varepsilon, SDK_R)$; When R is a special value,$Ugsc(m, SDK_S, VEK_R) = Ver(\tau, VEK_S)$; When S and R are both not special values, $Ugsc(w, SDK_R, VEK_S)$=$Usc(w, SDK_R, VEK_S)$.

Han proposed a Generalized Signcryption ECGSC based on ECDSA [4].Following is the scheme:

1) **Parameters**: Parameters of the elliptic curve
   - the parameters follow the SEC1 standard, which can be described as a sixtuple $T = (p, a, b, G, n, h)$;
   - G is a base point;
   - $ord(G) = n$;
   - O is the infinite element of group (G).
2) **Syntax**:In the scheme there are the syntax as following
   - $Q = [x]G$ denotes the scalar multiplex on the elliptic curve;
   - $\|$ denotes connecting two messages;
   - $\in_R$ denotes randomly choosing an element in one set;
   - $Bind$ denotes Alice and Bob's identity;
   - $\{0,1\}^l$denotes binary sequence of length l;
   - $K_{enc}, K_{mac}, K_{sig}$ is a binary sequence;
   - $H : \{0,1\}^* \rightarrow \mathbb{Z}_p^*$ and $K : \mathbb{Z}_p^* \rightarrow \{0,1\}^{\mathbb{Z}+*}$ denote two hash functions;
   - $LH(.) : \{0,1\}* \rightarrow \{0,1\}^{l+z}$ denotes hash function output long digest, we can choose $SHA-256$,$SHA-384$ or $SHA-512$;
   - $MAC_k : \{0,1\}^l \times \{0,1\}^t \times \{0,1\}^z$ denote message authenticate function which has key k. $|k| = t$, $|m| = l$,$l + |MAC(.)| = |LH(x_2)|$;
   - These hash functions have property :$H(0) \rightarrow 0$,$K(0) \rightarrow 0$,$LH(0) \rightarrow 0$,$MAC(0) \rightarrow 0$.
3) **Key generation**(n, T):Generate user's private and public key
   - Generate Alice's private and public key,choose $d_A \in_R \{1, \cdots, , n-1\}$,$Q_A = [d_A]G$,return (d_A, Q_A);
   - Generate Bob's private and public key,$d_B \in_R \{1, , n-1\}$,$Q_B = [d_B]G$,return (d_B, Q_A);
   - Generate null user's private and public key $(0, O) \leftarrow Gen(U, T)$,$U \in \Phi$.
4) **Generalized Signcryption** $SC(m, d_A, Q_B)$: it consists of seven algorithms
   - $k \in_R 1, \cdots, n-1$;
   - $R = [k]G = (x_1, y_1)$,$r = x_1 \bmod p$;
   - $[k]P_B = (x_2, y_2)$;
   - $K_{enc} = LH(x_2)$,$(K_{mac}, K_{sig}) = K(y_2)$;
   - If $d_A = 0$, $s = \phi$, Else $s = k_{-1}(H(m \parallel Bind \parallel K_{sig}) + rd_A) mod n$;
   - $e = MAC_{K_{mac}}(m)$;
   - $c = (m \parallel e) \oplus K_{enc}$,Return $w = (c, R, s)$.
5) **Generalized Unsigncryption** $DSC(w, d_B, Q_A)$: it also consists of seven algorithms
   - $r = x(R)$(R's x axiom);
   - $(x_2, y_2) = [d_B]R$;
   - $K_{enc} = LH(x_2)$, $(K_{mac}, K_{sig}) = K(y_2)$;
   - $(m||e) = c \oplus K_{enc}$;
   - $e' = MAC_{K_{mac}}(m)$, If $e \neq e'$,return $\perp$else if $s = \phi$,return m;
   - $u_1 = s^{-1}H(m||Bind||K_{sig})$,$u_2 = s^{-1}r$;
   - $R' = [u_1]G + [u_2]Q_A$;If $R' \neq R$, return $\perp$,else return m.

B. Security Notions for Generalized Signcryption

Because Generalized Signcryption can work as encryption, signature or signcryption schemes, the adversary can get more oracles' service. For example, when considering confidentiality of Generalized Signcryption in encryption-mode, we must note adversary can get both Decryption Oracle service and Unsigncryption Oracle service. Note that Unsigncryption Oracle can maybe help the adversary decrypt challenge ciphertext. Analogously, when considering unforgeability of Generalized Signcryption in signature-mode, we must note adversary can get Signature Oracle service and Signcryption Oracle service. When considering confidentiality of Generalized Signcryption in signcryption-mode, we must note that the adversary can get Unsigncryption Oracle service and Decryption Oracle service. When considering unforgeability of Generalized Signcryption in signcryption-mode, we must note adversary can get Signature Oracle service and Signcryption Oracle service.

When talking about attacking against encryption schemes, we always emphasis on Decryption Oracle, but in fact, there is also an Encryption Oracle. But because public key is known to all, every one can get this Oracle's service, and it does not give the adversary any more attacking power than usual user. So we often omit this Oracle. The same thing happens in signature and signcryption schemes. Actually for Generalized Signcryption scheme, the adversary can get six types of Oracle's services: Encryption Oracle, Decryption Oracle, Signature Oracle, Verifying Oracle, Signcryption Oracle and Usigncryption Oracle.

Definition 2: **(Confidentiality in Encryption-mode)** Given security parameter $k = |p|$,let

$Adv^{IND-CCA2}_{GSC^{ENC},A}(k)$=$Pr[Exp^{IND-CCA2-1}_{GSC^{ENC},A}(k) = 1]$ - $Pr[Exp^{IND-CCA2-0}_{GSC^{ENC},A}(k) = 1]$

For $b \in \{0, 1\}$,the following is the experiment:

Experiment $Exp^{ind-cca2-b}_{GSC^{ENC},A}(k)$
$pk_A, sk_A \leftarrow_R Gen(k, param)$;
$pk_B, sk_B \leftarrow_R Gen(k, param)$;
$(x_0, x_1, s) = A_1\{Enc_{pk_B}(.), Dec_{sk_B}(.), Sig_{sk_A}(.), Ver_{pk_A}(.),$
$Gsc_{sk_A,pk_B(.)}, Ugsc_{sk_B,pk_A}(.)\}(find)$;
$y = GSC^{ENC}_{pk_B}(x_b)$;
$d = A_2\{Enc_{pk_B}(.), Dec_{sk_B}(.), Sign_{sk_A}(.), Ver_{pk_A}(.),$
$Gsc_{sk_A,pk_B}(.), Ugsc_{sk_B,pk_A}(.)\}(x_0, x_1, y, s, guess)$;
Return d.

In the above attacking, A can get six services, the only restriction is that y cannot be queried to the Decryption Oracle $Dec_{sk_B}(.)$. If $Adv^{IND-CCA2}_{GSC^{ENC},A}(k)$is negligible, we say this Generalized Signcryption scheme is confidential when it work in encryption-mode.

Definition 3: **(Unforgeability in Signature-mode)** Given security parameter $k = |p|$, following is the experiment:

Experiment $ForgeExp^{cma}_{GSC^{SIG},F}(k)$
$pk_A, sk_A \leftarrow_R Gen(k, param)$;
$pk_B, sk_B \leftarrow_R Gen(k, param)$;
if$F^{Gsc_{sk_A,pk_B}(.),Ugsc_{sk_B,pk_A}(.)}_{Enc_{pk_B}(.),Dec_{sk_B}(.),Sig_{sk_A}(.),Ver_{pk_A}(.)}(.)$ output (m, s) which satisfy

- $Ver_{pk_A}(s) = T$;
- m has never been queried to $Sig_{sk_A}(.)$(existential unforgeable) or m is allowed to query $Sig_{sk_A}(.)$ but was never returned by $Sig_{sk_A}(.)$(strong unforgeable) ;

then return 1,else return 0.

In the above attacking, A can get six services, the only restriction is m has never been queried $Sig_{sk_A}(.)$(existential unforgeable) ,or m is allowed to query to $Sig_{sk_A}(.)$ but s was never returned by $Sig_{sk_A}(.)$(strong unforgeable). Let$Succ^{cma}_{Gsc^{SIG},F}(k) = Pr[Exp^{cma}_{GSC^{SIG},F}(k) = 1]$. If this value is negligible, we say this Generalized Signcryption scheme is unforgeable when it works in signature-mode.

Definition 4: **(Confidentially in Signcryption-mode)** Given security parameter $k = |p|$,let

$Adv^{IND-CCA2}_{GSC^{SC},A}(k)$=$Pr[Exp^{IND-CCA2-1}_{GSC^{SC},A}(k) = 1]$
-$Pr[Exp^{IND-CCA2-0}_{GSC^{SC},A}(k) = 1]$

For $b \in \{0, 1\}$,the following is the experiment:

Experiment $Exp^{ind-cca2-b}_{GSC^{SC},A}(k)$
$pk_A, sk_A \leftarrow_R Gen(k, param)$;
$pk_B, sk_B \leftarrow_R Gen(k, param)$;
$(x_0, x_1, s) = A_1\{Enc_{pk_B}(.), Dec_{sk_B}(.), Sig_{sk_A}(.), Ver_{pk_A}(.),$
$Gsc_{sk_A,pk_B(.)}, Ugsc_{sk_B,pk_A}(.)\}(find)$;
$c = GSC^{SC}_{pk_B,sk_A}(x_b)$;
$d = A_2\{Enc_{pk_B}(.), Dec_{sk_B}(.), Sign_{sk_A}(.), Ver_{pk_A}(.),$
$Gsc_{sk_A,pk_B}(.), Ugsc_{sk_B,pk_A}(.)\}(x_0, x_1, c, s, guess)$;
Return d.

In the above attacking, A can get six services, the only restriction is that c was never queried $Ugsc_{sk_B,pk_A}(.)$.If $Adv^{IND-CCA2}_{GSC^{SC},A}(k)$ is negligible, we say this Generalized Signcryption scheme is confidential when it works in signcryption mode.

Remark 1 What's the diffirence between Definition 2 and Definition 4? In definition 2,the challenge ciphertext cannot be queried to *Decryption Oracle*, but we can transform challenge ciphertext into some valid signcryption ciphertext and then query it to the *Unsigncryption Oracle*. In definiton 4, the challenge signcryption ciphertext cannot be queried to *Unsigncryption Oracle*,but we can transform the challenge signcryption ciphertext to some valid ciphertext and then query it to the *Decryption Oracle*.

Definition 5: **(Unforgeablity in Signcryption-mode)** Given security parameter $k = |p|$, following is the experiment:
Experiment $ForgeExp^{cma}_{GSC^{SC},F}(k)$
$pk_A, sk_A \leftarrow_R Gen(k, param)$; $pk_B, sk_B \leftarrow_R Gen(k, param)$;
if $F^{Gsc_{sk_A,pk_B}(.),Ugsc_{sk_B,pk_A}(.)}_{Enc_{pk_B}(.),Dec_{sk_B}(.),Sig_{sk_A}(.),Ver_{pk_A}(.)}(.)$ output (m, C) which satisfy
$-$ m has never been queried to $Gsc_{sk_A,pk_B}(.)$;

$- Ugsc_{sk_B,pk_A}(C) = m$;
then return 1,else return 0.
In the above attacking, A can get six services, the only restriction is that c was never returned by $Gsc_{sk_A,pk_B}(.)$. Let$Succ^{cma}_{GSC^{SC},F}(k) = Pr[Exp^{cma}_{GSC^{SC},F}(k) = 1]$. If this value is negligible, we say the Generalized Signcryption scheme is unforgeable when it works in signcryption-mode.

Remark 2 What's the diffirence between Definition 3 and Definition 5? In definition 3,the forged signature is not the output of *signature Oracle*,but can be the transformation of some valid result returned by *Signcryption Oracle*. In definiton 5, the forged signcryption ciphertext is not the output of *Signcryption Oraxle* but can be the transformation of some valid result returned by *Signature Oracle*.

III. An Improved Generalized Signcryption Based on ECDSA

A. An attack on this Scheme and Some Remarks

Attack In the ECGSC scheme the adversary intercept the ciphertext $w = (c, R, s)$set $s = \phi$, query the new ciphertext $w = (c, R, \phi)$ to *Decryption Oracle*, the *Decryption Oracle* will return m, which break the confidentiality of Generalized Signcryption in signcryption-mode. Note here, the adversary does not query $w = (c, R, s)$ to *Unsigncryption Oracle*, which is the only restriction for the adversary. The attack can be successful just because we use *Decryption Oracle* to decrypt the modified challenge signcryption ciphertext.

Remark 3 The origin scheme depend on hash function with additional property, that is, $H(0) \rightarrow 0$,$K(0) \rightarrow 0$,$LH(0) \rightarrow 0$,$MAC(0) \rightarrow 0$.But we know, if there exists non-change point in hash function, this would bring bad effects to the hash function. Especially, for hash function working in CBC mode, this can be damage. Another reason is that hash function with addition property can not be easily devised. It does not follow principal of modern hash family. So we suggest deleting this additional property.

Remark 4 The original scheme uses if/else clause, and the conditional variant is s ,and s is just a local variant, programs with normal access rights can modify it. For example, some adversary can just add some program in the origin scheme's code at proper time, let $s = \phi$, he would get the plaintext m. So we suggest delete the if-clause in the algorithm.

B. An Improved Generalized Signcryption Based on ECDSA

In this section, we give an improved Generalized Signcryption scheme. Improved scheme has the same parameter, syntax with the origin scheme. But we do not need hash function satisfy $H(0) \rightarrow 0$,$K(0) \rightarrow 0$,$LH(0) \rightarrow 0$, $MAC(0) \rightarrow 0$, and we introduce another point Q, which can be any point not belonging to the elliptic curve (or no one would choose this point as his public key).Here we can assume $Q = (0, 0)$. The reason we introduce this point is for encryption-mode and signature-mode. We define a function $f(t)$. if $t = Q$, $f(t) = 0$,if $t \neq Q$, then $f(t) = 1$. For signcryption-mode, $Bind = SH(Q_A||Q_B)$, for encryption-mode, $Bind = SH(Q_A||Q)$,for signature-mode, $Bind = SH(Q||Q_B)$.SH represents hash function, its output is 32 bit, and we denote its length by $|sh|$. We change the length of LH's output to $l + z + |sh|$, we denote $|K_{sig}| = |sig|$.

1) **Parameters**: Same as the original scheme.
2) **Syntax**:Almost same as the original scheme except we do not need hash functions with additional property,introduce a new point and modify some syntex's meaning.
   - we do not need hash function satisfy $H(0) \rightarrow 0$,$K(0) \rightarrow 0$,$LH(0) \rightarrow 0$, $MAC(0) \rightarrow 0$;
   - we introduce another point Q, which can be any point not belonging to the elliptic curve (or no one would choose this point as his public key).Here we can assume $Q = (0, 0)$. The reason we introduce this point is for benefitting encryption-mode and signature-mode. We define a function $f(t)$. if $t = Q$, $f(t) = 0$,if $t \neq Q$, then $f(t) = 1$;
   - SH represents hash function, its output is 32 bit, and we denote its length by $|sh|$. We change the length of LH's output to $l + z + |sh|$, we denote $|K_{sig}| = |sig|$;
   - For signcryption-mode, $Bind = SH(Q_A||Q_B)$, for encryption-mode, $Bind = SH(Q_A||Q)$,for signature-mode, $Bind = SH(Q||Q_B)$.
3) **Key generation**(n, T):Same as the original scheme.
4) **Generalized Signcryption** $SC(m, d_A, Q_A, Q_B)$: it consists of seven algorithms
   - Compute $f(Q_A)$,$f(Q_B)$,
   - $k \in_R 1, \cdots, n - 1$;
   - $R = [k]G = (x_1, y_1)$,$r = x_1 \bmod p$;
   - $[k]P_B = (x_2, y_2)$;
   - $K_{enc} = f(Q_B) * LH(x_2)$,$(K_{mac}, K_{sig}) = f(Q_B) * K(y_2)$;
   - If $d_A = 0$, $s = \phi$, Else $s = k_{-1}(f(Q_A) * H(m \parallel Bind \parallel K_{sig}) + f(Q_A) * rd_A)$ $modn$;
   - $e = f(Q_B) * MAC_{K_{mac}}(m)$;
   - $c = (m \parallel e) \oplus K_{enc}$,Return $w = (c, R, s)$.
5) **Generalized Unsigncryption** $DSC(w, d_B, Q_A, Q_B)$: it also consists of seven algorithms
   - Compute $f(Q_A)$,$f(Q_B)$,
   - $r = x(R)$(R's x axiom);
   - $(x_2, y_2) = [d_B]R$;
   - $K_{enc} = f(Q_B) * LLH(x_2), (K_{mac}, K_{sig}) = f(Q_B) * LK(y_2)$;
   - $(m||e) = c \oplus K_{enc}$;
   - $e' = f(Q_B) * LMAC_{K_{mac}}(m)$, If $e \neq e'$,return $\perp$else if $s = \phi$,return m;
   - $u_1 = s^{-1} * f(Q_A) * H(m||Bind||K_{sig})$,$u_2 = s^{-1} * f(Q_A) * r$;
   - $R' = [u_1]G + [u_2]Q_A$;If $R' \neq R$, return $\perp$,else return m.

IV. Security Proofs for Our Improved Generalized Signcryption

The idea of the origin scheme's security proofs is the following. When the Generalized Signcryption work as in signcryption-mode, the author can reduce confidentiality of signcryption to a scheme proposed by Krawczyk in Crypto 2001 [1], and this scheme is proved to be ciphetext unforgeable under chosen plaintext attacks. We denote this encryption scheme ATEOTP and the analog Elliptic Curve's variant ECATEOTP. But the author just discussed the Signcryption Oracle service, no caring about other Oracle service, this is not sufficient. [5] can also reduce SUF-CMA of signcryption to SUF-CMA of ECDSA, but the reduction is uncorrect.Also [5] do not give secrurity proof for generalized signcryption working in encryption-mode and signature- mode.This paper tries to solve these problems.

A. Prove SUF-CMA of the Generalized Signcryption in Signcryption-mode

We will apply a standard technique of provable security theory game hopping in our proofs. We define a sequence of games:$G0$,$G1$. they are reduced from the real attacking game . In every game, the private and public key, the adversary and the Random Oracle's coin flipping space are not changed. The difference comes from the view defined by rules. We will reduce the attack to SUF-CMA of ECGSC to SUF-CMA of ECDSA. Assume the success probability of attacking SUF-CMA is τ, its running time is T. We denote character with $*$ as the forged ciphetext and its related variables.

GAME G0: In $GAMEG0$, we just use the standard technique of simulating hash function. We can know this environment and the really environment is indistinguishable in the random oracle model. Let S_0 denote attacking successfully, assume $Pr[S_0] = \varepsilon$.

1) Simulate Random Oracle $LH(x)$Query $LH(x)$,if the record (x, lh) is found in LH-list, then Oracle return lh else randomly choose $lh \in \{0,1\}^{l+z+|sh|}$,add (x, lh) to the H-list;
2) Simulate Random Oracle $K(y)$:Query $K(y)$,if the record (y, k) is found in K-list, then Oracle return k ,else randomly choose $k \in \{0,1\}^{z+|sig|}$,add (y, k) to K-list.
3) Simulate Random Oracle H:Query $H(m \parallel SH(Q_A \parallel Q_B) \parallel K_{sig})$,if the record $(m \parallel SH(Q_A \parallel Q_B) \parallel K_{sig}, h)$ is found in H-list, then Oracle return h ,else randomly choose $h \in \{0,1\}^|p|$ add record $(m \parallel SH(Q_A \parallel Q_B) \parallel K_{sig}, h)$ to H-list.
4) Simulate Random Oracle MAC:Query $MAC(K_{mac}, m \parallel SH(Q_A \parallel Q_B) \parallel s)$,If the record $(K_{mac}, m \parallel SH(Q_A \parallel Q_B) \parallel s, mac)$ is found in MAC-list, then Oracle return mac ,else randomly choose $mac \in \{0,1\}^z$,add the record $(K_{mac}, m \parallel SH(Q_A \parallel Q_B) \parallel s, mac)$ into the MAC-list.
5) Simulate Signcryption Oracle Sc:Real Signcryption in real environment. In assume adversary can get this service.
6) Simulate Unsigncryption Oracle Usc:Think about insider adversary. Because the adversary know the receiver's private key, he can get this integrated service (The simulator just gives the receiver's private key to the adversary).
7) Simulate Encryption Oracle Enc:Because the adversary can get the Encryption Oracle service by only needing to know the receiver's public key, but this is public to all. So the adversary can get the integrated service. (The simulator just gives the receiver's public key to the adversary).
8) Simulate Decryption Oracle Dec:Think about insider adversary. Because the insider adversary know the receiver's private key, he can get the integrated service. (The simulator just gives the receiver's private key to the adversary).
9) Simulate Sign And Verify Oracle Sig/Ver:In this game, assume the adversary can get the integrated service of Sign Oracle. Because implementing Verify Oracle just needs the signer's public key, and the public key is known to all. So the adversary can get this integrated service.
10) How to forge valid signcryption ciphertext:Assume the forged ciphetext is $w^* = (c^*, R^*, s^*)$the only restriction is that w^* was not queried to Sc Oracle.Totally there are two methods of forging ciphetext: One is by attacking signcryption directly, the other is utilizing Sign Oracle. Note the adversary can forge new valid signcryption ciphetext by utilizing Sign Oracle.

GAME G1: In this game, we will remove the restriction of linkage of encryption and signature in simulating GSC Signcryption Oracle. We remove the layer of encryption and reduce signcryption scheme to ECDSA signature scheme. We will substitute Sign Oracle by ECDSA algorithm. Other oracles are simulated as in $GAMEG0$.

1) Simulate Signcryption Oracle Gsc
   - Add new elements of $(\diamondsuit, (K_{mac}, K_{sig}))$ in K-list. Note we must set the first item of new element vacant; we give it some value later. Add new elements of $(\diamondsuit, K_{enc})$ in H-list. We also set the first item of new element vacant, we will give it some value later.
   - Call algorithm of $ECDSA(m \parallel SH(Q_A \parallel Q_B) \parallel K_{sig}, d_A)$in Random Oracle, let$(m \parallel SH(Q_A \parallel Q_B) \parallel K_{sig}, R, s)$be the output result. In this process there will be a H-list;
   - Find element of $(K_{mac}, m \parallel SH(Q_A \parallel Q_B) \parallel s)$in MAC-list. If $(K_{mac}, m \parallel SH(Q_A \parallel Q_B) \parallel s, K_{mac})$ is found in the MAC-list, then we return mac. Else, choosing randomly $mac \in \{0,1\}^z$return mac,add record of $(K_{mac}, m \parallel SH(Q_A \parallel Q_B) \parallel s, mac)$in MAC-list;
   - Compute $c = (m \parallel SH(Q_A \parallel Q_B) \parallel mac) \oplus K_{enc}$;

- Let (c, R, s) be the output of Signcryption Oracle Gsc when the input is (m, d_A, Q_A, Q_B);

2) Now we think about how to map vacant of elements in K-list and H-list to (x_2, y_2) . Because the simulator know the private key, so it can decryption the ciphertext. First we show how to simulate the Unsigncryption Oracle, in this process, we can give this map
3) Simulate Unsigncryption Oracle Ugsc
   - Query (c, R, s) to Unsigncryption Oracle Ugsc;
   - The simulator compute $(x_2, y_2) = d_B R$;
   - First we find s in the second item of $(K_{mac}, m \parallel SH(Q_A \parallel Q_B) \parallel s, mac)$ MAC-list. If s is found in $(K_{mac}, m \parallel SH(Q_A \parallel Q_B) \parallel s, mac)$, return $K_{mac} m \parallel SH(Q_A \parallel Q_B) \parallel s, mac$ else return ”Invalid Ciphertext”;
   - Next find K_{mac} in the second item of elements in K-list. If K_{mac} is found in $(\diamondsuit, (K_{mac}, K_{sig}))$-list, let the first item of this element be y_2, else return ”Invalid Ciphertext”;
   - Compute $t = c \oplus m \parallel SH(Q_A \parallel Q_B) \parallel mac$ and find t in the LH-list. If t is found equal to some element of $(\diamondsuit, K_{enc})$, then let the first item of this element be x_2, else return ”Invalid Ciphertext”.
4) Simulate Sign Oracle Sig:Using algorithm of $ECDSA(m \parallel SH(Q_A \parallel Q)), d_A)$,
   let its output be Sign Oracle's output.

Remark 5:In the above simulation,we use a technique different from usual. Here we use the condition that attacker can know the receiver's private key and can compute $[d_B]R$ and x_2, y_2.So we can find the relationship between x_2, y_2 and (K_{mac}, K_{sig}),K_{enc}.

$GameG1$ and $GameG0$ are indistinguishable, except some queries have been given to k-list,LH-list before simulation or some ciphertexts have been guessed correctly by adversary. Assume the adversary has queried K-Random Oracle,H-Random Oracle,LH-Random Oracle,MAC-Random Oracle $q_K, q_H, q_{LH}, q_{MAC}$ times, denote S_1 as the adversary forges successfully in GAME G1, then

$$| Pr[S_0] - Pr[S_1] | \leq \frac{q_H}{2^{|p|}} + \frac{q_{LH}}{2^{l+z+|SH|}} - \frac{q_H}{2^{|p|}} \cdot \frac{q_{LH}}{2^{l+z+|SH|}} \cdot \frac{q_{MAC}}{2^z} \cdot \frac{q_K}{2^{z+|Sig|}}$$

Theorem 1: If the adversary A can forge valid signcryption ciphertext of Generalized Signcryption in signcryption-mode successfully with probability τ and the running time is T.Assume A queries K-Random Oracle,H-Random Oracle, LH-Random Oracle, MAC-Random Oracle $q_K, q_H, q_{LH}, q_{MAC}$ times, queries Signcryption Oracle, Sign Oracle, Encryption Oracle, Unsigncryption Oracle, Verify Oracle, Decryption Oracle $q_{Gsc}, q_{Ugsc}, q_{Sig}, q_{Ver}, q_{Enc}, q_{Dec}$ times. Then he forges signature of ECDSA with probability ϵ,

$$\epsilon \geq \tau - (\frac{q_H}{2^{|p|}} + \frac{q_{LH}}{2^{l+z+|SH|}} - \frac{q_H}{2^{|p|}} \cdot \frac{q_{LH}}{2^{l+z+|SH|}} \cdot \frac{q_{MAC}}{2^z} \cdot \frac{q_K}{2^{z+|Sig|}})$$

The running time

$$T' \geq T + (q_{LH} + q_K)f + (q_{Gsc} + q_{Sig})g$$

f denote the running time of computed$d_b R$ one time,g denote the running time of compute kG one time

B. Prove Confidentiality of the Generalized Signcryption in Signcryption-mode

We reduce confidentiality of the Generalized Signcryption in signcryption-mode to confidentiality of ECATEOTP which as following.

Definition 6: ECATEOTP is an encryption scheme, and we know it's IND-CCA2 secure [1].

1) Encryption $Enc(m, Q_A, Q_B)$
   - $k \in_R \{1, \cdots, n-1\}$;
   - $(x_1, y_1) = R = [k]G$
   - $(x_2, y_2) = [k]Q$;
   - $K_{enc} = LH(x2), (K_{mac}, K_{sig}) = K(y_2)$;
   - $e = MAC_{K_{mac}}(m \parallel SH(Q_A \parallel Q_B))$;
   - $c = (m \parallel SH(Q_A \parallel Q_B \parallel e) \oplus K_{enc}$;
   - Return $w = (c, R)$.
2) Decryption $Dec(w, d_B, Q_A, Q_B)$
   - $[d_B]R = (x_2, y_2)$;
   - $K_{enc} = LH(x2), (K_{mac}, K_{sig}) = K(y_2)$;
   - $(m \parallel SH(Q_A \parallel Q_B) \parallel e) = c \oplus K_{enc}$;
   - $e' = MAC_{K_{mac}}(m \parallel SH(Q_A \parallel Q_B))$;
   - if $e = e'$,return ” $\perp$ ”;else return m.

Assume the success probability of forging Valid Ciphertext of ECATEOTP is η , and running time is T.

GAME G0: In $GAMEG0$, we just use the standard technique of simulating hash function. We can know this environment and the really environment is indistinguishable in the random oracle model. Let S_0 denote attacking successfully, assume $Pr[S_0] = \gamma$.

1) Simulate Random Oracle $LH(x)$,$K(y)$,H,MAC: Same as common name oracles in section 4.1;
2) Simulate Signcryption Oracle Sc:Think about insider adversary. Because the adversary know the sender's private key, he can get this integrated service;
3) Simulate Unsigncryption Oracle Usc:Real Unsigncryption under real environment. Assume adversary can get this service;
4) Simulate Encryption Oracle Enc:The adversary can get the Encryption Oracle service by only needing to know the receiver's public key. And this is public to all, so the adversary can get this integrated service;
5) Simulate Decryption Oracle Dec:Assume the adversary can get this integrated service;
6) Simulate Sign And Verify Oracle Sig/Ver:Think about insider adversary. Because insider adversary know the

receiver's private key, he can get this integrated service.The adversary can get the Verify Oracle service by only needing to know the sender's public key, but this is public to all. So the adversary can get this integrated service.

7) How to decrypt challenge ciphertext:Denote the challenge ciphertext$(c*, R*, s*)$. There are two ways to decrypt the challenge ciphertext: One is to utilize attacking on the signcryption scheme. The other is to use Decryption Oracle.

GAME G1: In this game, we try to reduce Unsigncryption Oracle to Decryption Oracle of ECATEOTP and substitute Decryption Oracle of Generalized Signcryption by Decryption Oracle of ECATEOTP..

1) Simulate Signcryption Oracle Gsc
   - Everything is done honestly just as in the real Signcryption Algorithm. But when some queries to the Random Oracle LH, K, H, and MAC, we return something following the standard technique of simulating Hash Function.
2) Simulate Unsigncryption Oracle Ugsc
   - There have been LH, K, H, MAC-list in simulate Signcryption Oracle Gsc;
   - Using Decryption Oracle of ECATEOTP: $Dec(w, d_B, Q_A, Q_B)$in Random Oracle;
   - Algorithm Dec will compute $(x_2, y_2) = [d_B]R$,it must get value of $LH(x_2)K(y_2)$ according to LH-list, K-list. It finds (x_2, K_enc) and $(y_2, (K_{Mac}, K_{sig}))$ in K-list and LH-list. If the element is found, then return the second item of element; else return "Invalid Ciphertext";
   - Compute $(m \parallel Bind \parallel e) = c \oplus K_{enc}$;
   - Find $m \parallel SHQ_A \parallel Q_B \parallel K_{sig}$ in the first item of elements in H-List. If $(m \parallel SH(Q_A \parallel Q_B) \parallel K_{sig}, h)$ is found, Simulator return h.Else return "Invalid Ciphertext";.
   - Compute $u_1 = s^{-1} * hu_2 = s^{-1} * r$;
   - Compute $R' = [u_1]G + [u_2]Q_A$ If $R' \neq R$,return $\perp$else return m.
3) Simulate Decryption Oracle Dec:Using algorithm of $Dec(w, d_B, Q, Q_B)$, let its output be Decryption Oracle's output.

$GAMEG1$ and $GAMEG0$ are indistinguishable, except some ciphertexts have been guessed validly by adversary. Assume the adversary has queried K-Random Oracle, H-Random Oracle, LH-Random Oracle, MAC-Random Oracle q_K, q_H, q_{LH},
q_{MAC}times, denote S_1 as the adversary forges successfully in $GAMEG1$, then

$$|Pr[S_0] - Pr[S_1]| \leq \frac{q_H}{2^{|p|}} \cdot \frac{q_{LH}}{2^{l+z+|SH|}} \cdot \frac{q_{MAC}}{2^z} \cdot \frac{q_K}{2^{z+|Sig|}}$$

Theorem 2: If the adversary A can attack confidentiality of Generalized Signcryption in signcryption-mode successfully with probability η , the running time is T.Assume A queries K-Random Oracle, H-Random Oracle,LH-Random Oracle, MAC-Random Oracle times, queries Signcryption Oracle, Sign Oracle, Encryption Oracle, Unsigncryption Oracle, Verify Oracle, Decryption Oracle $q_{Gsc}, q_{Ugsc}, q_{Sig}, q_{Ver}, q_{Enc}, q_{Dec}$ times. Then he can attack IND-CCA2 property of ECATEOTP with probability

$$\zeta > \eta + \frac{q_H}{2^{|p|}} \cdot \frac{q_{LH}}{2^{l+z+|SH|}} \cdot \frac{q_{MAC}}{2^z} \cdot \frac{q_K}{2^{z+|Sig|}}$$

The running time

$$T' \geq T + (q_{LH} + q_K)f + (q_{Gsc} + q_{Sig} + q_{Ugsc}, q_{Ver}, q_{Enc}, q_{Dec})g$$

f denote the running time of computedd_bR one time,g denote the running time of compute kG one time

C. Prove SUF-CMA of the Generalized Signcryption in Sgnature-mode

When Generalized Signcryption Oracle work as a signature scheme, Generalized Signcryption is actually ECDSA. So we omit the proof and give the following theorem.

Theorem 3: If the adversary A can attack SUF-CMA of Generalized Signcryption in signature-mode successfully with probability η, the running time is T. Then he can forge valid signature of ECDSA with probability

$$\mu \approx \eta$$

The running time $T' = T$.

D. Prove Confidentiality of the Generalized Signcryption in Encryption-mode

When Generalized Signcryption Oracle work as an encryption scheme, Generalized Signcryption is actually ECATEOTP. So we omit the proof and give the following theorem.

Theorem 4: If the adversary A can attack confidentiality of Generalized Signcryption in encryption-mode successfully with probability η, and the running time is T. Then he can forge valid ciphertext of ECATEOTP with probability

$$\mu \approx \eta$$

The running time $T' \approx T$.

V. Conclusion and Open Problems

Based on Han et al's paper [3]–[5] our paper pay attention to the formal model of Generalized Signcryption. We give an improved Generalized Signcryption scheme based on ECDSA and give its security proof . We remark that this paper just gives a Generalized Signcryption scheme based on ECC, there are still much work can be done on this new primitive.So we propose following open problems to develop generalized signcryption research.

1) Give more experiments on the efficiency advantage over solely signcryption.
2) Propose more generalized signcryption schemes based on discrete logarithm problem.

3) Propose generalized signcryption schemes based on integer factoring problem.
4) Propose generalized signcryption schemes based on identity-based cryptography([21] has partially solved this question,but we can hope more).
5) Consider universal compose security for generalized signcryption. And this maybe be quite complicated for this cryptographic primitive can not lie in the current framework of universal composable security.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China under contract no. 60842006. The authors would like to express their gratitude thanks for Dr. Yiliang Han's many helpful discussions.

REFERENCES

[1] H. Krawczyk The order of encryption and authentication for protecting communications (or: How secure is SSL?). In *Advances in Cryptology, Proc. CRYPTO2001,* LNCS 2139, pages 310–331. Springer–Verlag, 2001.

[2] Y. Zheng Digital signcryption or how to achieve cost (signature &encryption) $\ll$ cost (signature) $+$ cost (encryption). In *Advances in Cryptology, Proc. CRYPTO 1997,* LNCS 1294, pages 165–179. Springer–Verlag, 1997.

[3] Y. Han, X. Yang. ECGSC: Elliptic Curve based Generalized Signcryption Scheme,Cryptology Eprint Archive, 2006/126.

[4] Y. Han, X. Yang. New ECDSA-Verifiable Generalized Signcryption. *Chinese Journal of Computer, No. 11.*, pages. 2003–2012, 2006.

[5] Y. Han. Generalization of Signcryption for Resources-constrained Environments. *Wireless Communication and Mobile Computing*, pages. 919–931, 2007.

[6] J. Malone-Lee, Mao W. Two birds one stone: Signcryption using RSA. In *Topics in Cryptology - CT-RSA 2003,* LNCS 2612, pages. 210–224. Springer–Verlag, 2003.

[7] Y. Zheng, H. Imai. How to construct efficient signcryption schemes on elliptic curves. *Information Processing Letters,, Vol. 68, No. 5, Sep.*, pages. 227–233, 1998.

[8] M. Bellare, C. Namprempre. Authenticated encryption: relations among notions and analysis of the generic composition paradigm. In *Advances in Cryptology, Proc. ASIACRYPT 2000,* LNCS 1976, pages 531–545. Springer–Verlag, 2000.

[9] J.H. An, Y. Dodis and T. Rabin On the security of joint signature and encryption.In *Advances in Cryptology, Proc. EUROCRYPT 2002,* LNCS 2332, pages 83–107. Springer–Verlag, 2002.

[10] Y. Dodis, M. Rreedman, S. Jarecki and S. Walfish, Optimal signcryption from any trapdoor permutation. Cryptology ePrint Archive, Report: 2004/020, 2004.

[11] Y. Dodis, M. Rreedman, S. Jarecki, and S. Walfish, Versatile padding schemes for joint signature and encryption. In *Proceedings of Eleventh ACM Conference on Computer and Communication Security (CCS2004)*, pages 196–205. IEEE Computer Society, 2004.

[12] D. Alexander. Hybrid Signcryption Schemes With Outsider Security.In *Proceedings of The 8th Information Security Conference (ISC 2005),* LNCS 4212, pages. 203–217, Springer–Verlag, 2005.

[13] D. Alexander. Hybrid Signcryption Schemes With Insider Security. In *Proceedings of Information Security and Privacy 2005) (ACISP 2005),* LNCS 4307, pages. 253–266, Springer–Verlag, 2005.

[14] M. Bellare, P. Rogaway. Random oracle are practical: a paradigm for designing efficient protocols.In *Proceeding of the First ACM Conference on Computer and Communication Security (CCS1993)*, pages.62–73, IEEE Computer Society, 1993.

[15] J. Baek, R. Steinfeld and Y. Zheng. Formal Proofs for the Security of Signcryption.In *Public Key Cryptography'02 (PKC 2002),* LNCS 2274, pages. 80–98, Springer–Verlag, 2002.

[16] J. Stern, D. Pointcheval, J. Malone-Lee and N.Smart. Flaws in Applying Proof Methodologies to Signature Schemes.In *Advances in Cryptology-Crypto'02 (CRYPTO 2002),* LNCS 2442, pages. 93–110, Springer–Verlag, 2002.

[17] M. Bellare and P. Rogaway. Optimal Asymmetric Encryption - How to Encrypt with RSA.In *(Eurocrypt'94),* LNCS 950, pages. 92–111, Springer–Verlag, 1995.

[18] M. Bellare and P. Rogaway. The Exact Security of Digital Signatures -How to Sign with RSA and Rabin. In *(Eurocrypt '96),* LNCS 1070, pages. 399–416, Springer–Verlag, 1996.

[19] J. Baek, R. Steinfeld and Y. Zheng, Formal Proofs for the Security of Signcryption.*Journal of Cryptology, Vol. 20, Issue 2*, pages. 203–235, 2007.

[20] X. Boyen. Multipurpose identity-based signcryption:a swiss army knife for identity-based cryptography.In *(Crypto03)*, pages. 382–398, Springer–Verlag, 2003.

[21] L. Sunder and K. Prashant. ID based generalized signcryption,Cryptology Eprint Archive, 2008/084.

Xu an Wang was born in Feb. 23th, 1981. He obtained his bachelor and master's degree in the Engneering College of Chinese Armed Police Force. Now he is a lecturer in the same college, his main research fields are cryptography and information security.

Xiaoyuan Yang was born in Jan. 2th, 1959. He obtained his bachelor and master's degree in the Xidian University. Now he is a Professor in the Engneering College of Chinese Armed Police Force, he has pubilished almost 30 papers on different conferences and journals, his main research fields are cryptography and information security.

Jindan Zhang was born in April. 29th, 1983. She obtained her bachelor's degree in the Xidian University and her master's degree in the Shanxi University of Technology and Science. Now she is a lecturer in the Xianyang Vocational Technical College, her main research fields are digital watermark, cryptography and information security.

Examining Consumers' Willingness to Buy in Chinese Online Market

Shouming Chen
School of Economics and Management, Tongji University, Shanghai, China
Email: schen@tongji.edu.cn

Jie Li
School of Economics and Management, Tongji University, Shanghai, China
Email: tjlijie@gmail.com

***Abstract*—This paper aims to examine consumers' willingness to buy from an e-commerce vendor in Chinese online market by using empirical modeling. Initially, we hypothesize that six factors influence consumers' willingness to buy significantly. Then we develop 5 structural equation models (SEM) integrated these factors to test our hypotheses with the data from Chinese online market. The results suggest that perceived reputation and perceived risk are significantly associated with the level of consumers' willingness to buy, while perceived size, perceived system assurance, and perceived privacy information protection are insignificant related with the level of consumers' willingness to buy. In addition, perceived system assurance mediates the relationship between ease of use and willingness to buy.**

***Index Terms*—e-commerce, SEM, willingness to buy**

I. Introduction

In latest two decade, Internet as an information technology makes an egregious development and became indispensable in today's world. The prosperity of internet boosts the advent and development of Electronic Commerce, and continues to fascinate practitioners and enterprisers alike to invest. More and more companies engage their business online. On the other hand, the prosperity of e-commerce denotes that consumers have more and better choices when they purchase online than before. However, we are puzzled by a question, why consumers choose you when they have a number of choices. This question is a complex issue concerning consumer behavior in e-commerce. We focus the answer on the willingness to buy (WTB), which are directly associated with consumer purchase decision [1]. In this paper, we narrowed our research object on Chinese online market.

II. Literature Reviews

Internet exerts an increasingly strong influence on society and people's life. It provides a new space for people to communicate, entertain, study and work and so on. Online shopping, as we called e-commerce usually, is another amazing utility of Internet. With the number of Internet users growing, e-commerce would also expand. According to China Internet Network Information Center (CNNIC) 2007 report, in china alone, the number of internet user reached 210 million, 73 million more than the number in 2006 [2]. The CNNIC 2008 report shows that the number of internet user in China reached 253 million in June 2008 [3], and ranked as the top country that has the largest number of internet user in the world.

Further, the online market becomes more profitable and appealing when the internet user base is exploding. The growth of interest and market in the Internet as a shopping and purchasing medium is fascinating for practitioners and enterprisers alike. If you get online, you will find that Internet have been filling with B2B, C2C and B2C websites. The flourish of e-commerce engenders e-commerce venders' desires to understand and analyze the process of consumer purchase decision. When come to this topic, venders would like to ask a question, why consumers choose me when they have a lot of choices. We focus the answer on WTB, which are directly associated with consumer purchase decision [1].

According to the theory of reasoned action (TRA) model [1][4], an individual's performance in a specific behavior is determined by his or her behavioral intentions, which themselves are jointly determined by individual attitudes and subjective norms [1][5]. An extended model of TRA, namely the theory of planned behavior (TPB) was derived by adding perceived behavioral control as a determinant of behavior [6][7]. TRA and TPB have been empirically validated, and both models are widely used for predicting or explaining cognitive and affective behavior using the belief –> attitude –> intention –> behavior relationship in social psychology. Hence, WTB from an online vender, the measurement of purchase intention, is effective and appropriate to predict and explain the consumer purchase decision or behavior.

However, there is few researches study the factors that directly influenced WTB. Many researchers study WTB by making the correlation between consumer trust and WTB [8][9][10]. These researchers focus on the issues of consumer trust, setting WTB as "by-product" of consumer trust.

doi:10.4304/jcp.5.5.815-824

Gefen focus the study on Amazon.com and found that increased degrees of trust in an e-commerce vendor will increase people’s intentions to purchase products on that vendor's website [8]. Gefen and Straub demonstrate that four dimensions of trust, integrity, predictability, ability and benevolence, influence the level of intention to purchase [9]. Proposing model of consumer trust in e-commerce vendors for the US, Singapore and China groups, Teo and Liu found that the positive relationships exist between consumer trust and their attitude toward a vendor and between consumers’ attitude and their willingness to buy from the vendor [10].

All of these studies use the similar reasoning approach: factors (influence on trust) —> consumer trust —> WTB. After analyzing the line of reasoning, we have to point out that the factors that influence consumer trust perhaps are not the same factors that influence WTB. It is probability that some factors that significantly influence consumer trust do not affect WTB at all. In addition, definitions of trust vary from one context to another. Tan and Sutherland find that the Oxford dictionary holds 17 definitions for the word trust, that the Webster’s holds 18 definitions for the word trust, and that more than 12 different definitions have been used in consumer trust researches [11]. Different definitions require different factors that affected trust [12] [13] [14] [15] [16]. Hence, it is inappropriate for an e-commerce vendor to evaluate or to improve his customers’ WTB by using the factors that affect consumer trust. Therefore, it is necessary and valuable to identify the factors that influenced WTB directly. These factors are more appropriate and valid to evaluate WTB than the factors that affect consumer trust are.

From previous consumer behavior researches, we could infer that some factors will affect consumers’ willingness to buy from an e-commerce vendor. These factors include perceived reputation (PREP), perceived risk (PR), ease of use (EOU), perceived size (PS), perceived system assurance (PSA), and perceived privacy information protection (PPIP). However, few researches examine or document the significance of relationship between these factors and willingness to buy, let alone such examination in Chinese online market. China is acknowledged to be a fast growing economic entity and keep to exert an increased influence on global economy. Subsequently, Chinese business market, including Chinese online market, attracts attentions and investments from all over the world. It is undoubted that many individuals and organizations interest in the factors influencing consumers’ willingness to buy that largely contributes the success of online business. Therefore, the purpose of this paper is to examine consumers’ willingness to buy from an e-commerce vendor in Chinese online market.

III. Conceptual Model and Research Hypotheses

In order to examine consumers’ willingness to buy in Chinese online market, a conceptual model was developed (shown in Fig. 1). The conceptual model shows the proposed hypotheses, the possible links between variables: perceived reputation (PREP), perceived risk (PR), ease of use (EOU), perceived size (PS), perceived system assurance (PSA), and perceived privacy information protection (PPIP). The proposed links are hypothesized in the following.

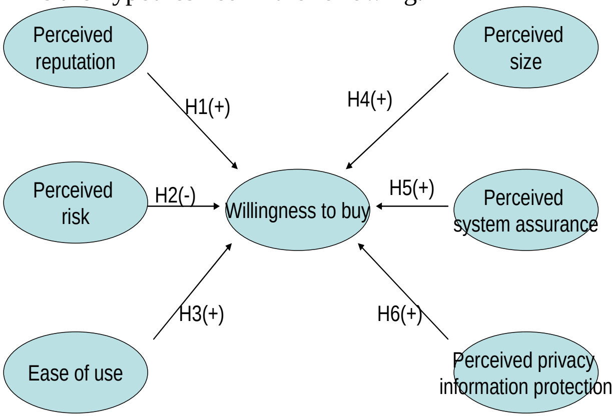


Figure 1. Conceptual model

A. *Perceived reputation (PREP)*

Reputation refers to the extent to which buyers believe a seller is professionally competent or honest and benevolent [12]. Researchers have recognized that a firm’s reputation is a valuable intangible asset that requires a long-term investment of resources, efforts, and attention to customer relationships [10]. In the traditional marketing literature, reputation has been shown to be positively related to the buyer’s positive judgment in the seller [17]. In Internet shopping, perceived reputation of a vendor has also been revealed to be significantly related to buyer’s positive judgment, such as trust, in the vendor [18] [19] [20]. Jarvenpaa et al. asserted that customers’ perceptions of an Internet store’s reputation affect their trust in the store [20]. All of these researches show that reputation have a positive judgment of an e-commerce vendor. Therefore, we postulate that:

H1. The perceived reputation of an e-commerce vendor is positively related to the level of willingness to buy.

B. *Perceived risk (PR)*

Unlike consumers in the physical market, consumers may be dealing with remote vendors that they have never met and products that cannot be touched and felt. Hence, consumers tend to be reluctant to conduct businesses based only on the information provided by e-commerce vendors because such information may not be reliable. Bauer argues that once a risk has been perceived in a purchase situation, there seems to be some reasonable evidence that subsequent consumer behavior is shaped by this risk perception [21]. Perceived risk could also be regarded as a belief about situations. Mayer et al. defined risk perception as the trustor’s belief about likelihoods of gains and losses outside of considerations that involve the relationships with the particular trustee [22]. Therefore, in accordance with TRA [1], consumer’s perceived risk might have a negative relationship with their WTB. Ruyter et al. empirically verified that perceived risk has an impact on consumers’ attitudes toward e-service [23]. McKnight et al. stated that trusting intention is likely to be fragile if the perceived risk is high [24]. Consequently, we assume that:

H2. The perceived risk is negatively associated with the level of willingness to buy.

C. *Ease of use (EOU)*

Derived from the technology acceptance model (TAM) that introduced and developed by Fred Davis [25], ease of use in website was considered to be one of influence elements in our model. TAM is a model derived from a theory that addresses the issue of how users come to accept and use a technology. Perceived ease of use is one of the main variables, which are hypothesized to be fundamental determinants of user acceptance [26]. Davis and Arbor define perceived ease of use as the degree to which a person believes that using a particular technology will be free of effort [26]. Users believe that a given application is useful, but they may, at the same time, believe that the technology is too hard to use and that the performance benefits of usage are outweighed by the effort of using the application [26]. Hence, we assume that:

H3. The ease of use in website is positively related to the level of willingness to buy.

D. *Perceived size (PS)*

Size refers to a seller's overall size and market share position [12]. Since a large market-share firm must serve a more diverse and heterogeneous set of customers [27], a large overall size and market share suggests that the firm consistently delivers on its promises to its consumers. Otherwise, it would not have been able to maintain its position in the industry [12]. Large organizational size also indicates that the firm is likely to possess expertise and necessary support systems that encourage trust and loyalty [28]. Larger firms also tend to have more well-developed Web sites to encourage transactions [29]. Finally, in an e-commerce environment, large size suggests that the vendor is able to assume the risk of product failure or transit losses and to compensate buyers accordingly [20]. All of these advantages brought by size make us to assume that:

H4. The perceived size of an e-commerce vendor is positively related to the level of willingness to buy.

E. *Perceived system assurance (PSA)*

System assurance is defined as the dependability and security of a vendor's online transaction system, which enables transactions through the Internet be secure and successful [10]. Consumers will perceive risks when they perceived that the web they interacting lack sufficient security of transaction system. Risks perceived reduce the level of WTB. Ambrose and Johnson found that insufficient trust in the security and reliability of the transactions over the Internet is a commonly expressed concern of consumers [30]. In addition, Kini and Choobineh argued that the assurance properties of the system that consumers interact with are critical in developing and maintaining consumer trust [31]. Teo and Liu found that consumer trust is significantly related to WTB [10]. Hence, it follows that:

H5. Perceived system assurance is positively related to the level of willingness to buy.

F. *Perceived privacy information protection (PPIP)*

It is clear that consumer concern with privacy of information has an impact on the consumer online market. Conducted by Business Week in 1998, a poll of 999 consumers revealed that privacy was the biggest obstacle preventing them from using Websites [32]. Another study by Forrester Research shows that two-thirds of consumers are worried about protecting personal information online [33]. What is more, a survey conducted by Harris in 2001 documented that consumer concerns about protecting its privacy on the Internet, as individuals who have not bought over the Internet list security of information storage and transmission and the use of personal information as the top reasons why they have not purchased [34]. All of these statistics shows that fears of privacy violations affect consumer willingness to buy online. Therefore, we postulate that:

H6. Perceived privacy information protection is positively associated with the level of willingness to buy.

IV. METHOD

In this part, we are going to introduce our research method, including data collection and measurements.

A. *Data collection*

Students in TONGJI University were used as subject in our study. Drennan et al. argued that university students are representative of a dominant cohort of online users [35]. The college students are experienced and regular users of the Internet, representing the most appropriate population of e-commerce user for e-commerce research.

In the survey, items of variables were developed by adapting existing measures to the research context. All items were scored on a five point Likert-type scale ranging from (1) Strongly Disagree to (5) Strongly Agree.

The survey questionnaire consisted of two sections. In the first section, respondents were asked to answer questions about basic information, such as the gender, education level, and so on. On the end of first section, we ask them to choose one of the five webs (taobao, joyo, ebay, dangdang and paipai) or to fill any other web from which they have purchased goods before. The second section consists of the questions measuring model variables. Respondents answered the questions in second section based on the web that they chose in the end of the first section.

We give out 300 survey questionnaires in TONGJI University Library. All of 300 questionnaires have been taken back. We rule out the questionnaires that were conducted incompletely and eliminate the questionnaires with no online purchase experiences. We screen out 256 valid questionnaires as our research sample. The basic information of respondents is shown in Table I.

TABLE I.
BASIC INFORMATION OF RESPONDENTS

ITEMS	FEATURES	NUMBER	PERCENTAGE
GENDER	Male	180	70.3%
	Female	76	29.7%
EDUCATION LEVEL	Bachelor	160	62.5%
	Master or MBA	83	32.4%
	PhD and above	13	5.1%
EXPENSE PER MONTH (yuan)	<1000	201	78.5%
	1000-3000	50	19.5%
	>3000	5	2.0%
ONLINE PURCHASE EXPERICENCE (years)	<1	15	5.9%
	1-2	113	44.1%
	2-3	89	34.8%
	>3	39	15.2%
ANNUAL ONLINE PURCHASE TIME	1	53	20.7%
	2-5	119	46.5%
	5-10	50	19.5%
	>10	34	13.3%
THE WEBSITE CHOSEN BY RESPONSER TO ANSWER THE SURVEY	Taobao	155	60.5%
	Joyo	44	17.2%
	Dangdang	38	14.8%
	Others	19	7.4%

B. Measurements

In this part, we will introduce the measurements in our research. The measurements contain seven latent variables and twenty three observable indicators. Each latent variable is measured by at least three observable indicators. All of the variables, indicators, descriptions are listed in TableⅡ.

In order to figure out the most valid descriptions for observable indicators in our questionnaire, all of the descriptions are tested and adjusted repeatedly to satisfy the requirements of reliability, validity and our research purpose. There are three types of description sources. The first type refers to the descriptions cited directly from previous researches, and we consider that they are already valid without adjusting. The second type refers to the descriptions cited from previous researches after adjusting. The third type refers to the new descriptions developed by ourselves with the reason that there are no related descriptions used before, such as the descriptions of PPIP. Although some descriptions are adjusted or new, all of them are derived from previous researches. Hence, we list the related references of the descriptions in Table Ⅱ.

TABLE II.
MEASUREMENTS

Latent Variable	Observable Indicator	Descriptions	References
Willingness To Buy (WTB)	WTB1	I prefer to shop in this website.	Jarvenpaa et al. [20]; Jarvenpaa and Tractinsky [19].
	WTB2	I would return to purchase from this website again.	
	WTB3	I like shopping in this website.	
Perceived Reputation (PREP)	PREP1	This website has a good reputation.	Jarvenpaa et al. [20]; Teo and Liu [10].
	PREP2	This website has a reputation for honest.	
	PREP3	This website has a reputation for fair.	
Perceived Risk (PR)	PR1	There is great uncertainty associated with purchasing online from this website.	Houghton et al. [36]; Teo and Liu [10].
	PR2	There is great purchasing risk of this website.	
	PR3	The vendor in this website may cheat consumer.	
	PR4	The good purchased is not the same as expected.	
Ease Of Use (EOU)	EOU1	It is not hard to use this website efficiently.	Davis et al. [25]; Davis and Arbor [26] .
	EOU2	It is convenient to find out the good I want to purchase.	
	EOU3	It is convenient to make a deal with the vendor in this website.	
Perceived Size (PS)	PS1	The size of this website is very large.	Jarvenpaa et al. [20]; Teo and Liu[10].
	PS2	This website is one of the biggest suppliers in industry.	
	PS3	This is a national or global website.	
Perceived System Assurance (PSA)	PSA1	The online transaction system of this website is stable.	Kini and Choobineh [31]; Teo and Liu [10].
	PSA2	It is safe to make a deal online from this website.	
	PSA3	This website has ability to accomplish online transaction successfully.	
Perceived Privacy Information Protection (PPIP)	PPIP1	This website will not give my privacy information away.	Branscum [33]; Harris [34].
	PPIP2	This website will not abuse my privacy information.	
	PPIP3	This website will protect my privacy information actively.	
	PPIP4	This website has ability to prevent my privacy information being stealing.	

V. RESULT

In this section, we use SPSS 13.0 to examine the reliability and validity and use LISREL 8.7 to conduct the Structural equation models.

A. Reliability

First, we focus on evaluating reliability. The Cronbach’s alphas of willingness to buy (WTB), perceived reputation (PREP), perceived risk (PR), ease of use (EOU), perceived size (PS), perceived system assurance (PSA), and perceived privacy information protection (PPIP) are 0.851, 0.875, 0.732, 0.723, 0.806, 0.761, and 0.870, respectively. All of these Cronbach’s alphas are >0.7. We conclude that all latent variables have adequate reliabilities.

B. models

According to the purpose of this paper, the latent variable, willingness to buy, is dependent variable, Y-variable, while the other six latent variables are independent variables, X-variables. After examining the reliability, we translate the conceptual model shown in Fig. 1 into an overall structural equation model, model 1, which integrated all six independent variables. The result shows that some independent variables significantly associate with the dependent variables while other independent variables do not. However, we could not make conclusion imprudently from overall model because some insignificant independent variables are likely to influence the significance of the relationship between other independent variables and dependent variable. Therefore, in order to make a precise and suggestive result, we develop another four structural equation models to get rid of the possible influence between variables. Table Ⅲ is a summery of the five models, including variables in each model, path coefficients, t-value and fix indices. Each X-variable contains two rows where the statistic number in the first row refers to its path coefficient and the statistic number bracketed in the second row refers to its t-value. The five structural equation models are demonstrated in the following.

TABLE III.
SUMMERY OF MODELS

Variables		MODELS				
		Model 1	Model 2	Model 3	Model 4	Model 5
Y-variable	Willingness to buy(WTB)	WTB	WTB	WTB	WTB	WTB
X-variable	Perceived reputation (PREP)	0.44* （4.99）	0.44* （5.42）	0.44* （4.97）	0.44* （5.38）	0.45* （5.45）
	Perceived risk (PR)	-0.20* （-2.82）	-0.20* （-2.94）	-0.20* （-2.90）	-0.20* （-2.92）	-0.20* （-2.86）
	Ease of use (EOU)	0.19 （1.61）	0.21* （2.38）	0.20* （2.12）	0.20 （1.70）	0.20* （2.14）
	Perceived size (PS)	0.01 （0.06）		0.01 （0.09）		
	Perceived system assurance (PSA)	0.01 （0.06）			0.01 （0.14）	
	Perceived privacy information protection (PPIP)	0.02 （0.25）				0.02 （0.36）
Fit indices	χ^2	419.32	109.25	166.68	186.69	229.87
	DF	209	59	94	94	109
	P-value	<0.001	<0.001	<0.001	<0.001	<0.001
	RMSEA	0.063	0.058	0.055	0.062	0.066
	CFI	0.96	0.97	0.97	0.97	0.98
	NFI	0.92	0.95	0.95	0.94	0.92
	NNFI	0.95	0.96	0.97	0.96	0.97
	GFI	0.87	0.94	0.92	0.92	0.90

Note: * means this coefficient is significant.

- Model 1

Model 1 is an overall structural equation model integrated all six independent variables (PREP, PR, EOU, PS, PSA and PPIP) to evaluate willingness to buy.

First, we test validity using PCA (Principle Component Analysis). The value of KMO is 0.828. The significance of Bartlett's Test is less than 0.001. KMO and Bartlett's Test indicates that all the factors are suitable for PCA. We use Varimax Rotation method to rotate factors. The rotated component matrix is shown in Table Ⅳ.

TABLE IV.
ROTATED COMPONENT MATRIX OF MODEL 1

	WTB	PREP	PR	EOU	PS	PSA	PPIP
WTB1	**0.837**	0.137	-0.096	0.137	0.145	0.019	0.092
WTB2	**0.825**	0.302	-0.057	0.119	0.068	0.054	0.093
WTB3	**0.789**	0.120	-0.244	0.142	0.127	0.144	0.057
PREP1	0.295	**0.793**	-0.092	0.055	0.242	0.160	0.125
PREP2	0.320	**0.747**	-0.038	0.212	0.275	0.195	0.008
PREP3	0.114	**0.773**	-0.136	0.248	0.166	0.166	0.081
PR1	-0.107	-0.067	**0.763**	0.109	0.004	-0.004	-0.086
PR2	-0.086	0.105	**0.738**	-0.054	0.008	-0.281	-0.138
PR3	-0.008	-0.277	**0.675**	-0.051	-0.030	0.169	0.017
PR4	-0.142	-0.018	**0.735**	-0.186	0.005	-0.031	-0.017
EOU1	0.122	0.274	-0.034	**0.756**	0.038	0.119	0.060
EOU2	0.074	0.103	-0.058	**0.636**	0.387	0.135	0.151
EOU3	0.211	0.070	-0.089	**0.750**	0.123	0.184	0.162
PS1	0.047	0.159	0.104	0.110	**0.824**	0.034	0.118
PS2	0.160	0.190	0.025	0.133	**0.817**	0.216	0.038
PS3	0.127	0.149	-0.128	0.117	**0.747**	0.145	0.028
PSA1	0.132	0.197	-0.031	0.284	0.182	**0.676**	0.216
PSA2	0.209	0.070	0.083	0.321	0.248	**0.648**	0.266
PSA3	-0.014	0.206	-0.117	0.043	0.111	**0.769**	0.156
PPIP1	0.012	0.068	0.041	0.060	0.055	0.230	**0.831**
PPIP2	-0.002	0.036	-0.086	-0.078	0.130	0.230	**0.859**
PPIP3	0.140	-0.055	-0.204	0.260	0.068	0.063	**0.730**
PPIP4	0.125	0.145	-0.030	0.170	-0.012	0.008	**0.859**

The rotated component matrix indicates that all the indicator items loaded very high (above 0.636) on their respective factors and below 0.40 all the other factors, suggesting good convergent validity and discriminant validity for each latent variable. Then we use Lisrel 8.7 to conduct SEM. The path diagram is shown in Fig 2. The fit indices of model 1 shown in Table Ⅲ indicate a good model fit. The standardized structural coefficients and t-value are shown in Table Ⅲ. Generally, when the absolute value of t>2, the coefficient is significant. Hence, the result of model 1 supports H1 and H2, but does not support H3, H4, H5 and H6.

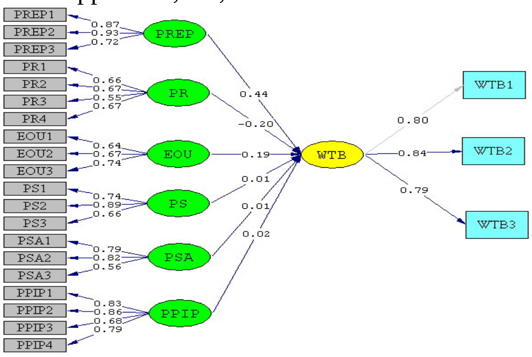


Figure 2. Path Diagram of Model 1

- Model 2

As we discussed before, model 1 did not rule out the possible influence between independent variables. Hence, we develop model 2 where three variables (PREP, PR and EOU) are integrated to evaluate WTB. First, we test validity using PCA (Principle Component Analysis). The value of KMO is 0.816. The significance of Bartlett's Test is less than 0.001. KMO and Bartlett's Test indicates that all the factors are suitable for PCA. We use Varimax Rotation method to rotate factors. The rotated component matrix is shown in Table Ⅴ.

TABLE V.
ROTATED COMPONENT MATRIX OF MODEL 2

	WTB	PREP	PR	EOU
WTB1	**0.846**	0.178	-0.087	0.145
WTB2	**0.823**	0.319	-0.068	0.110
WTB3	**0.804**	0.141	-0.234	0.226
PREP1	0.289	**0.851**	-0.094	0.127
PREP2	0.307	**0.815**	-0.028	0.273
PREP3	0.104	**0.803**	-0.141	0.274
PR1	-0.098	-0.071	**0.774**	0.070
PR2	-0.086	0.051	**0.754**	-0.151
PR3	-0.024	-0.243	**0.669**	0.056
PR4	-0.141	-0.013	**0.736**	-0.181
EOU1	0.132	0.244	-0.026	**0.733**
EOU2	0.087	0.197	-0.044	**0.751**
EOU3	0.208	0.100	-0.098	**0.801**

The rotated component matrix indicates that all the indicator items loaded highly (above 0.669) on their respective factors and below 0.40 all the other factors, suggesting good convergent validity and discriminant validity for each latent variable. Then we use Lisrel 8.7 to conduct SEM. The path diagram is shown in Fig 3. The fit indices of model 2 shown in Table Ⅲ indicate a good model fit. The standardized structural coefficients and t-value are shown in Table Ⅲ. According to the t-value, the result of model 2 supports H1, H2 and H3.

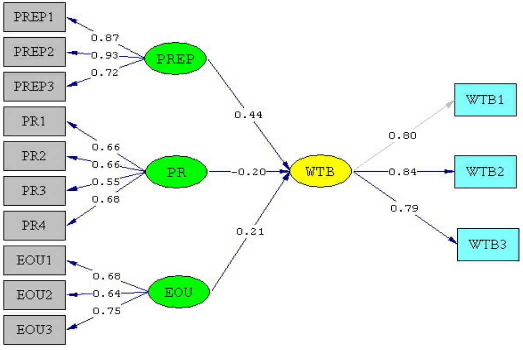


Figure 3. Path Diagram of Model 2

- Model 3

In model 3, we add model 2 with another X-variable, perceived size (PS), to evaluate WTB. The value of KMO is 0.831. The significance of Bartlett's Test is less than 0.001. KMO and Bartlett's Test indicates that all the factors are suitable for PCA. We use Varimax Rotation to rotate factors. We use Varimax Rotation method to rotate factors. The rotated component matrix is shown in Table Ⅵ.

TABLE VI.
ROTATED COMPONENT MATRIX OF MODEL 3

	WTB	PREP	PR	EOU	PS
WTB1	**0.843**	0.150	-0.092	0.122	0.146
WTB2	**0.822**	0.324	-0.064	0.112	0.066
WTB3	**0.800**	0.114	-0.242	0.206	0.143
PREP1	0.289	**0.809**	-0.101	0.085	0.259
PREP2	0.305	**0.774**	-0.037	0.234	0.283
PREP3	0.107	**0.789**	-0.143	0.256	0.180
PR1	-0.104	-0.056	**0.778**	0.082	-0.009
PR2	-0.082	0.063	**0.757**	-0.136	-0.043
PR3	-0.016	-0.269	**0.660**	0.046	0.018
PR4	-0.141	-0.025	**0.734**	-0.187	0.016
EOU1	0.126	0.292	-0.017	**0.766**	0.029
EOU2	0.082	0.103	-0.066	**0.665**	0.398
EOU3	0.207	0.095	-0.100	**0.792**	0.141
PS1	0.050	0.168	0.099	0.089	**0.830**
PS2	0.153	0.222	0.017	0.153	**0.828**
PS3	0.134	0.152	-0.121	0.158	**0.749**

The rotated component matrix indicates that all the indicator items loaded very high (above 0.660) on their respective factors and below 0.40 all the other factors, suggesting good convergent validity and discriminant validity for each latent variable. Then we use Lisrel 8.7 to conduct SEM. The path diagram is shown in Fig 4. The fit indices of model 3 shown in Table Ⅲ indicate a good model fit. The standardized structural coefficients and t-value are shown in Table Ⅲ. According to the t-value, the result of model 3 supports H1, H2 and H3 but does not support H4.

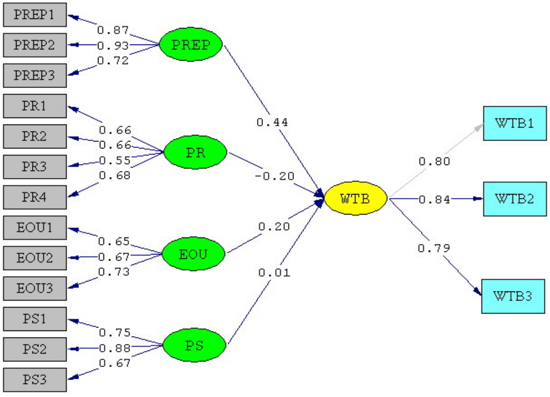


Figure 4. Path Diagram of Model 3

- Model 4

Model 4 was also developed from model 2 by adding another X-variable, perceived system assurance (PSA). The value of KMO is 0.822. The significance of Bartlett's Test is less than 0.001. KMO and Bartlett's Test indicates that all the factors are suitable for PCA. We use Varimax Rotation to rotate factors. The rotated component matrix is shown in Table Ⅶ.

TABLE VII.
ROTATED COMPONENT MATRIX OF MODEL 4

	WTB	PREP	PR	EOU	PSA
WTB1	**0.842**	0.171	-0.088	0.146	0.067
WTB2	**0.827**	0.302	-0.061	0.128	0.070
WTB3	**0.794**	0.135	-0.241	0.176	0.139
PREP1	0.299	**0.824**	-0.084	0.089	0.212
PREP2	0.317	**0.790**	-0.020	0.240	0.200
PREP3	0.112	**0.781**	-0.133	0.258	0.193
PR1	-0.113	-0.061	**0.770**	0.104	-0.045
PR2	-0.090	0.097	**0.743**	-0.030	-0.320
PR3	-0.010	-0.284	**0.674**	-0.065	0.182
PR4	-0.139	-0.021	**0.738**	-0.212	0.013
EOU1	0.121	0.252	-0.036	**0.752**	0.083
EOU2	0.091	0.166	-0.043	**0.701**	0.241
EOU3	0.216	0.064	-0.094	**0.766**	0.222
PSA1	0.135	0.215	-0.035	0.287	**0.734**
PSA2	0.220	0.099	0.077	0.356	**0.731**
PSA3	-0.018	0.222	-0.120	0.057	**0.761**

The rotated component matrix indicates that all the indicator items loaded very high (above 0.674) on their respective factors and below 0.40 all the other factors, suggesting good convergent validity and discriminant validity for each latent variable. Then we use Lisrel 8.7 to conduct SEM. The path diagram is shown in Fig 5. The fit indices of model 4 shown in Table Ⅲ indicate a good model fit. The standardized structural coefficients and t-value are shown in Table Ⅲ. According to the t-value, the result of model 4 supports H1 and H2 but does not support H3 and H5.

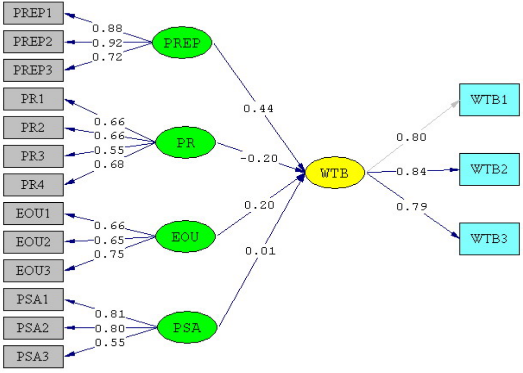


Figure 5. Path Diagram of Model 4

- Model 5

Model 5 was also developed from model 2 by adding another X-variable, perceived privacy information protection (PPIP). The value of KMO is 0.810. The significance of Bartlett's Test is less than 0.001. KMO and Bartlett's Test indicates that all the factors are suitable for PCA. We use Varimax Rotation to rotate factors. The rotated component matrix is shown in Table Ⅷ.

TABLE VIII.
ROTATED COMPONENT MATRIX OF MODEL 5

	WTB	PREP	PR	EOU	PPIP
WTB1	**0.848**	0.163	-0.089	0.147	0.087
WTB2	**0.821**	0.313	-0.067	0.105	0.086
WTB3	**0.797**	0.155	-0.228	0.203	0.080
PREP1	0.284	**0.849**	-0.093	0.105	0.142
PREP2	0.309	**0.813**	-0.036	0.274	0.035
PREP3	0.104	**0.798**	-0.143	0.269	0.092
PR1	-0.093	-0.078	**0.763**	0.099	-0.088
PR2	-0.080	0.063	**0.745**	-0.117	-0.189
PR3	-0.032	-0.225	**0.681**	0.012	0.066
PR4	-0.138	-0.016	**0.737**	-0.184	-0.015
EOU1	0.136	0.232	-0.033	**0.752**	0.056
EOU2	0.081	0.204	-0.035	**0.721**	0.165
EOU3	0.203	0.100	-0.088	**0.778**	0.176
PPIP1	0.016	0.095	0.044	0.105	**0.859**
PPIP2	0.004	0.087	-0.080	-0.019	**0.895**
PPIP3	0.152	-0.054	-0.205	0.262	**0.722**
PPIP4	0.124	0.119	-0.035	0.148	**0.835**

The rotated component matrix indicates that all the indicator items loaded very high (above 0.681) on their respective factors and below 0.40 all the other factors, suggesting good convergent validity and discriminant validity for each latent variable. Then we use Lisrel 8.7 to conduct SEM. The path diagram is shown in Fig 6. The fit indices of model 5 shown in Table Ⅲ indicate a good model fit. The standardized structural coefficients and t-value are shown in Table Ⅲ. According to the t-value, the result of model 5 supports H1, H2 and H3 but does not support H6.

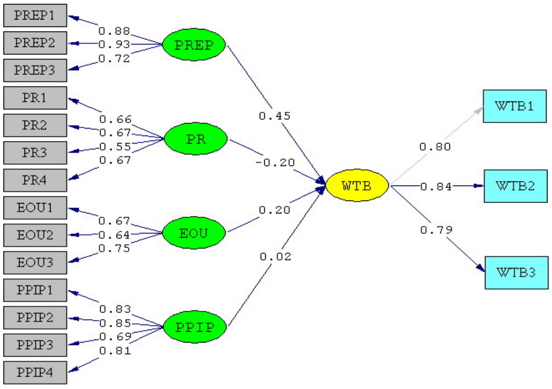


Figure 6. Path Diagram of Model 5

VI. DISCUSSION

This paper develops five structural equation models to demonstrate the hypotheses we proposed. The result supports some of them and denies the others.

H1 refers that the perceived reputation (PREP) of an e-commerce vendor is positively related to the level of willingness to buy (WTB). All of five models support H1. Accordingly, perceived reputation has a significantly positive relationship with WTB. This result is consistent with the findings of Teo and Liu [10] and Jarvenpaa et al. [20]. Both researches show that perceived reputation affects WTB. In e-commerce world, this result also predicts that higher vendor's reputation improve consumers' willingness to buy from the vendor.

H2 refers that the perceived risk (PR) is negatively associated with the level of willingness to buy (WTB). All five models support this hypothesis. The result provides empirical evidence to TRA [1] (Ajzen and Fishbein, 1980). Hence, high perceived risk will reduce the consumers' willingness to buy.

H3 refers that the ease of use (EOU) in website is positively related to the level of willingness to buy (WTB). Model 1 and model 3 both involving perceived system assurance (PSA) variable do not support H3. The models that do not involve perceived system assurance (PSA), such as model 2, model 3 and model 5, show that EOU is significantly link with WTB. Hence, we find that PSA is a modulator in the relationship between EOU and WTB. Here, we try to give possible explanations. In accordance with TAM, EOU is an effective factor that affects WTB. People increase the willingness to buy when they perceived that the website is ease to use. But, if they consider ease of use and system assurance together, they realize that system assurance play a more important role than the ease of use in shopping. Therefore, PSA affect the significance of the relationship between EOU and WTB, as shown in model 1 and model 3.

H4 refers that the perceived size (PS) of an e-commerce vendor is positively related to the level of willingness to buy (WTB). Model 3 shows that H4 is insignificant. This result is consistent with the findings of Teo and Liu [10] and Jarvenpaa et al. [20]. One possible reason is that size of an online vendor is less easily perceived on the Web than size of a physical store. Unlike size of a vendor in physical world, size of a vendor in the online world cannot be easily and correctly judged through its website. Hence, consumers may not care much about the size of an ecommerce vendor.

H5 refers that perceived system assurance (PSA) is positively related to the level of willingness to buy (WTB). Model 4 shows that H5 is insignificant. However, Teo and Liu argue that PSA significantly affects consumers' trust and then affects willingness to buy [10]. In addition, we find that PSA affect the significance of the relationship between EOU and WTB when we discuss H3, it is possible that PSA is correlate with EOU. Therefore, we could not conclude hastily that the relationship between PSA and WTB is insignificant. Further researches are required if we get a more precise conclusion.

H6 refers that perceived privacy information protection (PPIP) is positively associated with the level of willingness to buy (WTB). Model 5 shows that H6 is insignificant. This result is not consistent with the findings of Green et al. [32] and Branscum [33]. Although foreign survey and study reveal that privacy is an element influence consumers' willingness to buy. Perhaps Chinese people do not perceive privacy as important as foreign people do. The reason for such difference may derive from social and cultural background between two countries. We ask some respondents face to face that answer our questionnaire in TONGJI University Library. When come to privacy information protection, some of them feel that most e-commerce vendors will not protect customers' privacy information actively. They said that they care about privacy information protection, but it is impossible to prohibit their privacy information from being leaked. Accordingly, PPIP seems not to affect consumers' willingness to buy significantly. However, we would not suggest the conclusion that PPIP do not affect WTB. We need further researches to examine H6.

VII. Conclusion

This paper focuses on examining consumers' willingness to buy in Chinese online market. We review the previous literatures and pick out six variables, including perceived reputation, perceived risk, ease of use, perceived size, perceived system assurance, and perceived privacy information protection. Then we develop 5 models to test our hypotheses. The main conclusions from this paper are shown in the following. Perceived reputation of an e-commerce vendor is positively related to the level of willingness to buy. Perceived risk is negatively associated with the level of willingness to buy. However, perceived size, perceived system assurance, and perceived privacy information protection are insignificant related with the level of consumers' willingness to buy. Additionally, the relationship between ease of use and willingness to buy is mediated by perceived system assurance.

References

[1] I. Ajzen, M. Fishbein, Understanding Attitudes and Predicting Social Behavior, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[2] CNNIC 2007 report, http://www.cnnic.net.cn/uploadfiles/pdf/2008/1/17/104156.pdf

[3] CNNIC 2008 report, http://www.cnnic.net.cn/uploadfiles/pdf/2008/7/23/170516.pdf

[4] Fishbein M, Ajzen I. Belief, Attitude, Intention and Behaviour: An Introduction to Theory and Research. Addison-Wesley, 1975.

[5] Shih, H., An empirical study on predicting user acceptance of e-shopping on the Web, Information & Management, Vol. 41 No.3, 2004, pp.351-68

[6] I. Ajzen, From intentions to actions: a theory of planned behavior, in: J. Kuhl, J. Beckmann (Eds.), Action Control: From Cognition to Behavior, Springer, New York, 1985, pp. 11–39.

[7] I. Ajzen, Attitude structure and behavior, in: A.R. Pratkanis, S.J. Breckler, A.G. Greenwald (Eds.), Attitude Structure and Function, Lawrence Erlbaum, Hillsdale, NJ, 1989, pp. 241–274.

[8] Gefen D. E-commerce: the role of familiarity and trust. Omega 2000, 28(6):725–37.

[9] Gefen D, Straub DW. Consumer trust in B2C e-Commerce and the importance of social presence: experiments in e-Products and e-Services. Omega 2004; 32(6):407–24.

[10] Teo. T., Liu, J. Consumer trust in e-commerce in the United States, Singapore and China, Omega, Vol. 35 No.1, 2007, pp.22-38.

[11] Tan, F.B., Sutherland, P., Online consumer trust: a multi-dimensional model, Journal of Electronic Commerce in Organizations, Vol. 2 No.3, 2004, pp.40-58.

[12] Doney PM, Cannon JP. An examination of the nature of trust in buyer–seller relationships. Journal of Marketing 1997, 61 (2): 35–51.

[13] D M Rousseau, S B Sitkin, R S Burt,C Camerer.Not so different After all: A cross-Discipline View of Trust [J].The Academy of management Review,1998, 23(3), 393-404.

[14] Cheung, C., Lee, M., Trust in internet shopping: instrument development and validation through classical and modern approaches. Journal of Global Information Management, 9(3), 2001, 23-35.

[15] Gefen D., Reflections on the Dimensions of Trust and Trustworthiness among Online Consumers, The DATA BASE for Advances in Information Systems (33:3), 2002, pp. 38-53.

[16] Naquin, C.E., Paulson, G.D., Online bargaining and interpersonal trust, Journal of Applied Psychology, Vol. 88 No.1, 2003, pp.113-20.

[17] Ganesan S. Determinants of long-term orientation in buyer–seller relationships. Journal of Marketing 1994; 58(2): 1–19.

[18] McKnight DH, Choudhury V, Kacmar C. Trust in e-commerce vendors: a two-stage model. In: Proceedings of international conference on information systems (ICIS2000), Australia, December, 2000.

[19] Jarvenpaa SL, Tractinsky N. Consumer trust in an internet store: a cross-cultural validation. Journal of Computer Mediated Communication 1999; 5(2):1–35 Available from: http://www.ascusc.org/jcmc/vol5/issue2/jarvenpaa.html.

[20] Jarvenpaa SL, Tractinsky N, Vitale M. Consumer trust in an internet store. Information Technology and Management 2000; 1(1–2):45–71.

[21] Bauer RA. Consumer behavior as risk taking. In: Cox DF, editor. Risk taking and information handling in consumer behavior. Boston, MA: Harvard University Press; 1967. p. 23–33.

[22] Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. Academy of Management Review 1995, 20(3):709–34.

[23] Ruyter KD, Wetzels M, Kleijnen M. Customer adoption of eservice: an experimental study. International Journal of Service Industry Management 2001, 12(2): 184–207.

[24] McKnight DH, Cummings L, Chervany N. Initial trust formation in new organizational relationship. Academy of Management Review 1998; 23 (3): 473–90.

[25] Davis, Fred D., Bagozzi, Richard P. and Warshaw, Paul. User acceptance of computer technology a comparison of two theoretical models. Management Science Vo. 35 no. 8, 1989.

[26] Davis D. Fred, and Arbor, Ann., Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly September 1989.

[27] Griffin A, Hauser JR. The voice of the customer. Marketing Science 1993, 12(1):1–27.

[28] Chow S, Holden R. Toward an understanding of loyalty: the moderating role of trust. Journal of Managerial Issues 1997; 9(3): 275–98.

[29] Teo TSH, Pian Y. A model for web adoption. Information and Management 2003; 41(4):457–68.

[30] Ambrose PJ, Johnson GJ. A trust based model of buying behavior in electronic retailing. In: Proceedings of the

fourth Americas conference on information systems, Baltimore, MA, August, 1998.

[31] Kini A, Choobineh J. Trust in electronic commerce: definition and theoretical considerations. In: Proceedings of the 31st Annual Hawaii International Conference on System Sciences 4. Los Alamitos, CA: IEEE Computer Society Press; 1998. p. 51–61.

[32] Green H., Yang C., Judge P.C., 1998. A little privacy, please. Business Week 3569, 98–99.

[33] Branscum D., 2000. Guarding on-line privacy. Newsweek 135 (23), 77–78.

[34] Harris Interactive, Consumer privacy attitudes and beh aviors survey wave II, The Privacy Leadership Initiati ve, July 2001, http://www.understandingprivacy.org/content/library/harris2-execsum.pdf.

[35] J. Drennan, G.S. Mort, J. Previte, Privacy, risk perception, and expert online behavior: an exploratory study of household end users, Journal of Organizational and End User Computing 18, 2006, pp. 1–22.

[36] Houghton SM, Simon M, Aquino K, Goldberg C. No safety in numbers: the effects of cognitive biases on risk perception at the team level. Group and Organization Management 2000; 25(4):325–53.

Shouming Chen is an associate professor at School of Economics and Management, Tongji University. He received his Ph.D. in management from Fudan University in 2001. His research interests include strategic management and service management.

Jie Li was born in 1983. He received B.S degree from Tongji University in 2004. Currently, he is a graduate student at School of Economics and Management in Tongji University. He specializes in marketing management.

Call for Papers and Special Issues

Aims and Scope.

Journal of Computers (JCP, ISSN 1796-203X) is a scholarly peer-reviewed international scientific journal published monthly for researchers, developers, technical managers, and educators in the computer field. It provide a high profile, leading edge forum for academic researchers, industrial professionals, engineers, consultants, managers, educators and policy makers working in the field to contribute and disseminate innovative new work on all the areas of computers.

JCP invites original, previously unpublished, research, survey and tutorial papers, plus case studies and short research notes, on both applied and theoretical aspects of computers. These areas include, but are not limited to, the following:

- Computer Organizations and Architectures
- Operating Systems, Software Systems, and Communication Protocols
- Real-time Systems, Embedded Systems, and Distributed Systems
- Digital Devices, Computer Components, and Interconnection Networks
- Specification, Design, Prototyping, and Testing Methods and Tools
- Artificial Intelligence, Algorithms, Computational Science
- Performance, Fault Tolerance, Reliability, Security, and Testability
- Case Studies and Experimental and Theoretical Evaluations
- New and Important Applications and Trends

Special Issue Guidelines

Special issues feature specifically aimed and targeted topics of interest contributed by authors responding to a particular Call for Papers or by invitation, edited by guest editor(s). We encourage you to submit proposals for creating special issues in areas that are of interest to the Journal. Preference will be given to proposals that cover some unique aspect of the technology and ones that include subjects that are timely and useful to the readers of the Journal. A Special Issue is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

The following information should be included as part of the proposal:

- Proposed title for the Special Issue
- Description of the topic area to be focused upon and justification
- Review process for the selection and rejection of papers.
- Name, contact, position, affiliation, and biography of the Guest Editor(s)
- List of potential reviewers
- Potential authors to the issue
- Tentative time-table for the call for papers and reviews

If a proposal is accepted, the guest editor will be responsible for:

- Preparing the “Call for Papers” to be included on the Journal’s Web site.
- Distribution of the Call for Papers broadly to various mailing lists and sites.
- Getting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Instructions for Authors.
- Providing us the completed and approved final versions of the papers formatted in the Journal’s style, together with all authors’ contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

Special Issue for a Conference/Workshop

A special issue for a Conference/Workshop is usually released in association with the committee members of the Conference/Workshop like general chairs and/or program chairs who are appointed as the Guest Editors of the Special Issue. Special Issue for a Conference/Workshop is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

Guest Editors are involved in the following steps in guest-editing a Special Issue based on a Conference/Workshop:

- Selecting a Title for the Special Issue, e.g. “Special Issue: Selected Best Papers of XYZ Conference”.
- Sending us a formal “Letter of Intent” for the Special Issue.
- Creating a “Call for Papers” for the Special Issue, posting it on the conference web site, and publicizing it to the conference attendees. Information about the Journal and Academy Publisher can be included in the Call for Papers.
- Establishing criteria for paper selection/rejections. The papers can be nominated based on multiple criteria, e.g. rank in review process plus the evaluation from the Session Chairs and the feedback from the Conference attendees.
- Selecting and inviting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Author Instructions. Usually, the Proceedings manuscripts should be expanded and enhanced.
- Providing us the completed and approved final versions of the papers formatted in the Journal’s style, together with all authors’ contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

More information is available on the web site at http://www.academypublisher.com/jcp/.

(Contents Continued from Back Cover)