THE IEEE
Intelligent Informatics
BULLETIN

IEEE COMPUTER SOCIETY

IEEE Computer Society
Technical Committee
on Intelligent Informatics

June 2005 Vol. 5 No. 1 (ISSN 1727-5997)

Feature Articles

The Predicting Power of Textual Information on Financial Markets . . . . . *Gabriel Pui Cheong Fung, Jeffrey Xu Yu, & Hongjun Lu* 1
An On-Line Web Visualization System with Filtering and Clustering Graph . . . . . *Wei Lai, Xiaodi Huang, Ronald Wibowo, & Jiro Tanaka* 11
Association-Based Segmentation for Chinese-Crossed Query Expansion . . *Chengqi Zhang, Zhenxing Qin, & Xiaowei Yan* 18
A Partial-Repeatability Approach to Data Mining . . . . . *Kai-Yuan Cai, Yunfei Yin, & Shichao Zhang* 26
Web-based Multi-Criteria Group Decision Support System with Linguistic Term Processing Function . . . . . *Jie Lu, Guangquan Zhang, & Fengjie Wu* 35

Announcements

Related Conferences, Call For Papers/Participants 44

On-line version: http://www.comp.hkbu.edu.hk/~iib (ISSN 1727-6004)

The IEEE Intelligent Informatics Bulletin

Aims and Scope

The IEEE Intelligent Informatics Bulletin is the official publication of the Technical Committee on Intelligent Informatics (TCII) of the IEEE Computer Society, which is published twice a year in both hardcopies and electronic copies. The contents of the Bulletin include (but may not be limited to):

1) Letters and Communications of the TCII Executive Committee
2) Feature Articles
3) R & D Profiles (R & D organizations, interview profiles on individuals, and projects etc.)
4) Book Reviews
5) News, Reports, and Announcements (TCII sponsored or important/related activities)

Materials suitable for publication at the IEEE Intelligent Informatics Bulletin should be sent directly to the Associate Editors of respective sections.

Technical or survey articles are subject to peer reviews, and their scope may include the theories, methods, tools, techniques, systems, and experiences for/in developing and applying biologically and linguistically motivated computational paradigms, such as artificial neural networks, fuzzy logic, evolutionary optimization, rough sets, and self-organization in the research and application domains, such as data mining, Web intelligence, intelligent agent technology, parallel and distributed information processing, and virtual reality.

Publisher: *The IEEE Computer Society Technical Committee on Intelligent Informatics*
Address: *Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (Attention: Prof. Jiming Liu; Email: jiming@comp.hkbu.edu.hk)*
ISSN Number: *1727-5997 (printed) 1727-6004 (on-line)*
Abstracting and Indexing: *All the published articles will be submitted to the following on-line search engines and bibliographies databases for indexing — Google (www.google.com), The ResearchIndex (citeseer.nj.nec.com), The Collection of Computer Science Bibliographies (liinwww.ira.uka.de/bibliography/index.html), and DBLP Computer Science Bibliography (www.informatik.uni-trier.de/∼ley/db/index.html).*

The Predicting Power of Textual Information on Financial Markets

Gabriel Pui Cheong Fung†, Jeffrey Xu Yu†, Hongjun Lu‡

***Abstract*— Mining textual documents and time series concurrently, such as predicting the movements of stock prices based on the contents of the news stories, is an emerging topic in data mining community. Previous researches have shown that there is a strong relationship between the time when the news stories are released and the time when the stock prices fluctuate. In this paper, we propose a systematic framework for predicting the tertiary movements of stock prices by analyzing the impacts of the news stories on the stocks. To be more specific, we investigate the immediate impacts of news stories on the stocks based on the Efficient Markets Hypothesis. Several data mining and text mining techniques are used in a novel way. Extensive experiments using real-life data are conducted, and encouraging results are obtained.**

I. Introduction

In the financial markets, the movements of the prices are the consequences of the actions taken by the investors on how they perceive the events surrounding them as well as the financial markets. Investors' decisions on bidding, asking or holding the securities are greatly influenced by what others said and did within the financial markets [1], [2]. The emotions of fear, greed, coupled with subjective perceptions and evaluations of the economic conditions and their own psychological predispositions and personalities, are the major elements that affect the financial markets' behaviors [3], [4].

Yet, human behaviors are not random. People's actions in the financial markets, although occasionally irrational, are predominantly understandable and rational with respect to the social structure, social organization, perceptions and collective beliefs of this complex arena [1], [2], [5], [6]. At times, collective movements are launched which are in turn based on a group beliefs about how the markets will act or react [3], [4]. In these instances, trends develop which can be recognized, identified and anticipated to continue for some periods.

Nowadays, an increasing amount of crucial and valuable information highly related to the financial markets is widely available on the Internet.[1] However, most of these information are in textual format, such as news stories, companies' reports and experts' recommendations. Hence, extracting valuable information and figuring out the relationship between the extracted information and the financial markets are neither trivial nor simple.

†Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, {pcfung, yu}@se.cuhk.edu.hk

‡Department of Computer Science, The Hong Kong University of Science and Technology, luhj@cs.ust.hk

[1]E.g. Wall Street Journal: www.wsj.com, Financial Times: www.ft.com, Reuters: www.reuters.com, Bloomberg: www.bloomberg.com, CNN: www.cnnfn.com, etc

In this paper, we investigate how to utilize the rich textual information in predicting the financial markets. In contrast to the traditional time series analysis, where predictions are made based solely on the technical data (historical movements of the time series) and/or the fundamental data (the companies' profiles), this paper focuses on the problem of predicting the impacts of the textual information on the financial markets. To be more specific, real-time news stories and intra-day stock prices are used to denote respectively the information obtained from textual documents and the the movements of the financial markets. In other words, predictions are made according to the contents of news stories, rather than using the numerical data. This kind of problem is sometimes known as mining time series and textual documents concurrently [7], [8], [9], which is an emerging topic in the data mining community nowadays [10], [11], [12], [13]. The main contributions of this paper are summarized below:

1) **Figuring out the tertiary movements of the stock prices** A tertiary movement lasts for less than three weeks, which denotes the short-term stock market behavior [14], [15], [16]. In other words, tertiary movement is highly affected by the events surrounding the financial market. A new piecewise linear approximation algorithm, called *t-test based split and merge segmentation algorithm*, is proposed for figuring out the tertiary movements automatically.
2) **Detecting the relationship between the events mentioned in the news stories and the tertiary movements of stock prices** We have to correctly *align* and *select* the news stories to the tertiary movements of the stock price such that the aligned news stories are most likely to trigger or support the movements of the trends. The alignment process is based on the *Efficient Market Hypothesis* [1], [2] and the selection of the useful news stories is based on a χ^2 *estimation on the keywords distribution over the entire document collection.*
3) **Predicting the impact of a newly released news story on the stock price** Three kinds of impact are defined: `positive`, `negative` and `natural`. A piece of news story is said to have `positive` (or `negative`) impact if the stock price rises (or drops) significantly for a period after the news story is released; otherwise, if the the stock price does not fluctuate even after the news story is released, we said that the impact of the news story is `natural`. The major learning and prediction process is based on the text classification algorithm, *Support Vector Machines* [17].

The rest of this paper is organized as follows. Section II reviews the major preliminaries related to the discussion of this paper. Section III presents our proposed system. Section IV evaluates various aspects related to our proposed approach. A summary and conclusion is given in Section VI.

II. Preliminaries

The first systematic examination against the impacts of textual information on the financial markets is conducted by Klein and Prestbo [18]. Their survey consists primarily of a comparison of the movements of Dow Jones Industrial Average[2] with general news during the period from 1966 to 1972. The news stories that they have taken into consideration are the stories appearing in the "What's New" section in the *Wall Street Journal*, as well as three featured stories[3] carried on the Journal's front page. The major criticism of their study is that too few news stories are taken into consideration in each day. It is rather simple to assume that stories carried on the front page of the Wall Street Journal are enough for summarizing and reflecting the information appear in the whole newspaper. Interestingly, even with such a simple setting, Klein and Prestbo found that the pattern of directional correspondence, whether upwards or downwards, between the flow of the news stories and stock price movements manifested itself 80% of the time. Their findings strongly suggest that news stories and financial markets tend to move together.

Fawcett and Provost [10] formulate an *activity monitoring task* for predicting the stock price movements based on the content of the news stories. Activity monitor task is defined as the problem that involves monitoring the behaviors of a large population of entities for interesting events which require actions. The objective of the activity monitoring task is to issue alarms accurately and quickly. In the stock price movements detection, news stories and stock prices for approximately 6,000 companies over three months period are archived. An interesting event is defined to be a 10% change in stock price which can be triggered by the content of the news stories. The goal is to minimize the number of false alarms and to maximum the number of correctly predicted price spikes. It is worth noting that, the authors only provide a framework for formulating this predicting problem. The implementation details and an in-depth analysis are both missing. Perhaps this is because their main focus is not on examining the possibility of detecting stock price movements based on news stories, but is on outlining a general framework for formulating and evaluating the problems which require continuous monitoring their performance.

Thomas and Sycara [12] predict the stock prices by integrating the textual information that are downloaded from the web bulletin boards[4] into trading rules. The trading rules are derived by genetic algorithms based on numerical data. For the textual data, a maximum entropy text classification approach [19] is used for classifying the impacts[5] of the posted messages on the stock prices. For the trading rules, they are constructed by genetic algorithms based on the trading volumes of the stocks concerned, as well as the number of messages and words posted on the web bulletin boards per day. A simple market simulation is conducted. The authors reported that the profits obtained increased up to 30% by integrating the two approaches rather than using either of them. However, no analysis on their results is given.

Wuthrich et al. [13] develop an online system for predicting the opening prices of five stock indices[6] by analyzing the contents of the electronic stories downloaded from the Wall Street Journal. The analysis is done as follows: for each story, keywords are extracted and weights are assigned to them according to their significance in the corresponding piece of news story and on the corresponding day. By combining the weights of the keywords and the historical closing prices of a particular index, some probabilistic rules are generated using the approach proposed by Wuthrich [20], [21]. Based on these probabilistic rules, predictions on at least 0.5% price changes are made. The weaknesses of their system is that only the opening prices of financial markets could be predicted. Some others more challenging and interesting issues, such as intra-day stock price predictions, could not be achieved.

Following the techniques proposed by Wuthrick et al., Permunetilleke and Wong [11] repeat the work but with different a domain. News headlines (instead of news contents) are used to forecast the intra-day currency exchange rate (instead of the opening prices of stock indices). These news headlines belong to world financial markets, political or general economic news. They show that on a publicly available commercial data set, the system produces results are significantly better than random prediction.

Lavrenko et al. [9] propose a system for predicting the intra-day stock price movements by analyzing the contents of the real-time news stories. Analyst is developed based on a language modeling approach proposed by Ponte and Croft [22]. While a detailed architecture and a fruitful discussion are both presented in their paper, the following questions are unanswered: The authors claim that there should be a period, t, to denote the time for the market to absorb any new information (news stories) release, where t is defined as five hours. We have to admit that the market may spent time to digest information. However, such a long period may contradicts with most economic theories [1], [2]. In addition, news stories may frequently classify to trigger both the rise and drop movements of the stock prices in the training stage, which is a dilemma. Finally, in their evaluations, the impact of the news stories are "immediate" (without 5 hours time lag). This contradicts to the training phase of the system.

III. The Proposed System

In this paper, we are interested in determining whether a news story would have any impacts on the stock prices, and

[2] A financial index which composed of 30 blue-chip stocks listed on the New York Stock Exchange.

[3] Klein and Prestbo did not describe in details how they selected these three stories among all stories carried on the Journal's front page.

[4] Thomas and Sycara chose Forty discussion boards from `www.ragingbull.com`

[5] Two impacts are defined in their paper: up and down.

[6] These five stock indices are: Dow Jones Industrial Average, Nikkei 225, Financial Times 100 Index, Hang Seng Index and Singapore Straits Index

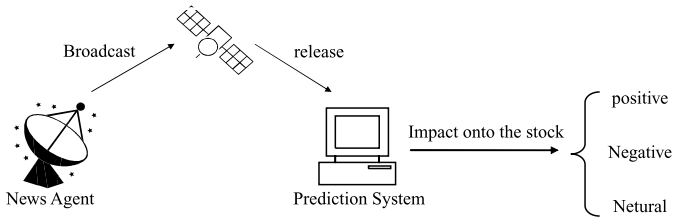


Fig. 1. The operation of the proposed system. After we received a news story from a news agent, we determine which of the three impact it has: `positive`, `negative` or `neutral`. A news story is said to have `positive impact` (or `negative impact`) if the stock price rise (or drop) signifi cantly for a period after the news story is released. If the stock price does not change after the news story is released, then the news story is regarded as `neutral`.

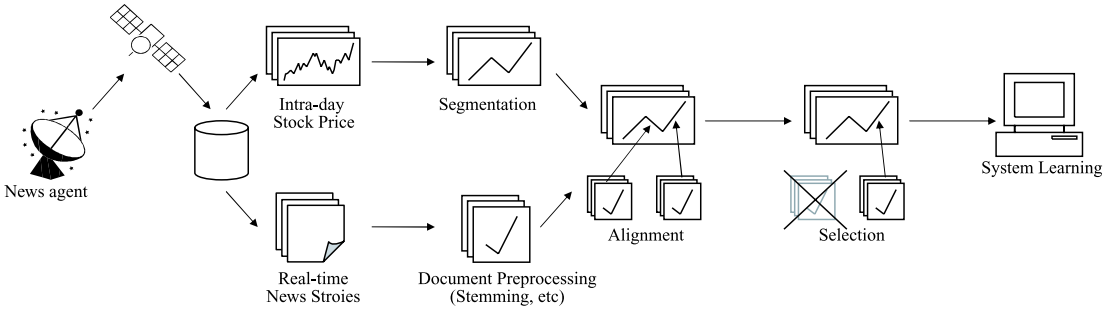


Fig. 2. The architecture of the proposed system. Four major processes are defi ned: 1) News Stories Alignment; 2) Time Series Segmentation; 3) Time Series Segmentation and 4) System Learning.

if so, what kinds of *impact* is this news story. Three impacts are defined: `positive`, `negative` and `neutral`. A news story is said to have `positive impact` (or `negative impact`) if the stock price rise (or drop) significantly for a period, T, after the news story has been broadcasted. If the stock price does not change after the news story is broadcasted, then the news story is regarded as `neutral`. Figure 1 illustrates the motivation described here.

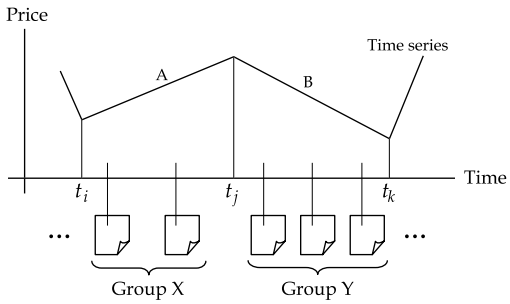
Figure 2 shows the architecture of the proposed system. For any prediction system to operate successfully, we first archive and label some sets of data and present them to the system for learning their relationships. These data are known as training data. The training data that we have taken are real-time news stories and intra-day stock prices. Since there are too many news stories and stocks in the market, such that it is impossible for us to read through the news stories one by one and classify their impact manually, we therefore must have a heuristics for selecting them automatically. We explain the details of the system in the following sections.

A. News Stories Alignment

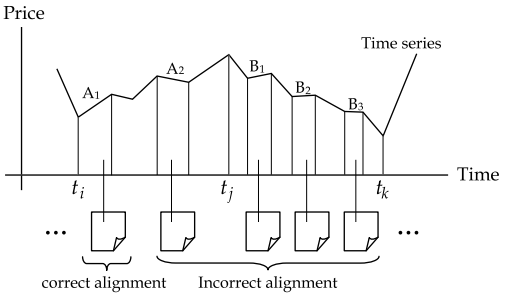
In order to obtain a set of reliable training data, we have to correctly align the news stories to the stock trend such that the aligned news stories is believed to trigger or support the movements of the trends. For aligning news stories to the stock time series, there could be three different formulation under different assumptions. They are further explained below:

1) **Formulation 1 – Observable Time Lag** In this formulation, there is a time lag between the news story is broadcasted and the stock price moves. It assumes that the stock market needs a long time for absorbing the new information. Let us take Figure 3 (a) to illustrate this idea. In this formulation, the Group X (news stories), is responsible for triggering Trend B, while Group Y does nothing with the two trends. Some reported works used this representation [9], [11], [12].
2) **Formulation 2 – Effi cient Market** In this formulation, the stock price moves as soon as after the new story is released. No time lag is observed. This formulation assumes that the market is efficient and no arbitrary opportunity normally exists. To illustrate this idea, let us refer to Figure 3 (a). Under this formulation, Group X is responsible for triggering Trend A, whild Group Y is responsible for triggering Trend B.
3) **Formulation 3 – Reporting** In this formulation, new stories are released only after the stock price has moved. This formulation assumes that the stock price movements are neither affected nor determined by any new information. The information (e.g. news stories) are only useful for *reporting* the situation but not *predicting* the future. Again, let us use Figure 3 (a) to illustrate this idea. Under this formulation, Group Y is responsible for accounting why Trend A would happened. Group X does nothing with the two trends.

Different scholars may in favor of one of the formulation. It is difficult, if not impossible, for finding a completely consensus. In this paper, we take the second formulation (Formulation 2 – Efficient Market), which is based on the Efficient Market Hypothesis. Thanks to Efficient Market Hypothesis, which states that the current market is an efficient information processor, such that it reflects the assimilation of



(a) An ideal situation.



(b) The truth situation.

Fig. 3. The alignment process. On the left: for the ideal situation, stories broadcast under the time series should be responsible for the fluctuation of the time series in that period. On the right: in reality, stock time series exhibit a high level of noise such that the impact of most stories are determined incorrectly.

all of the information available *immediately* [23], [24], we therefore align the news stories to the time series using the second formulation.

More formally, let d_i be a news story; $\mathcal{D}$ denote all of the news stories archived; $\mathcal{D}_{S_k}$ denote the documents that are aligned to segment S_k; $t_{rel}(d_i)$ denote the timestamp when the document d_i is released; $t_{begin}(S_k)$ and $t_{end}(S_k)$ denote the timestamp of segment S_k begin and the timestamp of segment S_k end, respectively. According to the second formulation that that documents which are broadcasted within a segment are aligned back to that segment:

$$d_i \in \{\mathcal{D}_{S_k} \mid t_{rel}(d_i) \geq t_{begin}(S_k) \text{ and } t_{rel}(d_i) < t_{end}(S_k)\} \quad (1)$$

However, no matter which formulation we take, note that all stock time series contain a high level of noise. Since every stock time series contains a high level of noise, such that even though the general trend is rising (or dropping), some dropping (or rising) segments can be observed. If we simply align the news stories based on the type of the time series segments (rise or drop), wrong alignment must be resulted. Figure 3 (b) illustrates this idea. In Figure 3 (b), even though the general trends from t_i to t_j is rising, the segment A_2 is slightly dropping. The news story releases under segment A_2 should be regarded as having `positive impact` (general trend) rather than having `negative impact` (exact observation). Similar situation is observed from t_j to t_k.

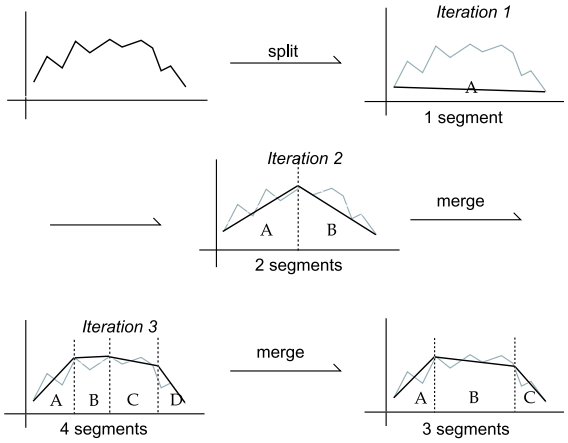


Fig. 4. General idea of the t-test based split and merge segmentation algorithm. The splitting phase aims at discovering all of the possible trends on the time series, while the merging phase aims at avoiding over-segmentation.

Algorithm 1 segment(T) – segment a time series T

1: T_{tmp} = split($T[t_0, t_n]$);
2: T_{final} = merge(T_{tmp});
3: **return** T_{final}

In order to remedy the above phenomenon, a higher level re-describing the time series into trends is necessary, e.g. re-describing Figure 3 (b) to Figure 3 (a) is necessary. This process is also known as time series segmentation. We provide our segmentation algorithm in the next section.

B. Time Series Segmentation

As with most data mining problems, data representation is one of the major elements to reach an efficient and effective solution. Since all stock time series contains a high level of noise, a high level time series segmentation is necessary for recognizing the trends on the times series. A sound time series representation involves issues such as recognizing the significant movements or detecting any abnormal behaviors, so as to study and understand its underlying structure.

Piecewise linear segmentation, or sometimes called piecewise linear approximation, is one of the most widely used technique for time series segmentation, especially for the financial time series [9], [25], [26]. It refers to the idea of representing a time series of length n using K straight lines, where $K \ll n$ [27], [28]. Most studies in this area are pioneered by Pavlidis et al. [28] as well as Dedua and Harts [29].

In this paper, we propose a t-test based split-and-merge piecewise linear approximation algorithm. The splitting phase aims at discovering trends on the time series, while the merging phase aims at avoiding over-segmentation. Figure 4 and Algorithm 1 illustrate the general idea of the proposed segmentation algorithm.4

t-test Based Split and Merge Segmentation Algorithm – Splitting phase

Let $T = \{(t_0, p_0), (t_1, p_1), \ldots, (t_n, p_n)\}$ be a financial time series of length n, where p_i is the price at time t_i for $i \in [0, n]$.

Initially, the whole time series is regarded as a single large segment, and is represented by a straight line joining the first and the last data points of the time series (Figure 4). In order to decide whether this straight line (segment) can represent the general trend of the time series, a one tail *t*-test is formulated:

$$\begin{aligned} H_0 &: \varepsilon = 0 \\ H_1 &: \varepsilon > 0 \end{aligned} \tag{2}$$

where ε is the expected mean square error of the straight line with respect to the actual fluctuation on the time series:

$$\varepsilon = \frac{1}{k} \cdot \sum_{i=0}^{k} (p_i - \hat{p}_i)^2 \tag{3}$$

where k is the total number of data points within the segment, $\hat{p}_i$ is the projected price of p_i at time t_i. The required t-statistics is:

$$t = \frac{\varepsilon}{\sqrt{\hat{\sigma}^2/n}} \tag{4}$$

where $\hat{\sigma}$ is the standard deviation of the mean square error, ε. The t-statistics is therefore compared with the t-distribution with $n-1$ degree of freedom using $\alpha = 0.05$. In other words, there is a probability of 0.05 that the null hypothesis ($H_0 : \varepsilon = 0$ in Equation (2), would be accepted given that it is incorrect.

The motivation of this formulation is that if the null hypothesis ($H_0 : \varepsilon = 0$) in Equation (2) is accepted, then the mean square error between the actual data points and the projected data points should be very small. Thus, the straight line, which is formulated by joining the first and the last data points of the segment, should be well enough to represent the trends of the data points in that segment. In contrast, if the alternative hypothesis is accepted ($H_1 : \varepsilon > 0$), then a single straight line is not well enough to represent the trend of the data points in the corresponding segment.

Let us consider for the case where the null hypothesis is rejected. If the null hypothesis is rejected, then the straight line is split at the point where the error norm is maximum, i.e. $\max_i \{(p_i - \hat{p}_i)^2\}$, and the whole process will be executed recursively on each segment (Figure 4 (b) – (c)). Algorithm 2 outlines the procedure of the splitting phase.

Algorithm 2 split($T[t_a, t_b]$) – split a time series T of length n from time t_a to time t_b where $0 \le a < b \le n$

```
1: T_temp = ∅
2: ε_min = ∞;
3: ε_total = 0;
4: for i = a to b do
5:    ε_i = (p_i − p̂_i)^2;
6:    if ε_min > ε_i then
7:       ε_min = ε_i;
8:       t_k = t_i;
9:    end if
10:   ε_total = ε_total + ε_i;
11: end for
12: ε = ε_total/(t_b − t_a);
13: if t-test.reject(ε) then
14:    T_temp = T_temp ∪ split(T[t_a, t_k]);
15:    T_temp = T_temp ∪ split(T[t_k, t_b]);
16: end if
17: return T_temp;
```

Algorithm 3 merge(T) – attempt to merge two adjacent segments on the time series T

```
1: while true do
2:    ε_min = ∞;
3:    repeat
4:       i = 0
5:       ε_i = Σ_{j=t'_i}^{t'_{i+2}} (p_j − p̂_j)^2;
6:       if ε_min > ε_i then
7:          ε_min = ε_i;
8:          k = i + 1;
9:       end if
10:   until end of the time series
11:   if t-test.accept(ε_min) then
12:      drop (t_k, p_k);
13:   else
14:      break;
15:   end if
16: end while
17: return T
```

t-test Based Split and Merge Segmentation Algorithm – Merging phase

After the splitting phase, *over-segmentation* will frequently occur. Over-segmentation refers to the situation where there exist two adjacent segments such that their slopes are similar, and they should be merged to from a single large segment. Let us refer to Figure 4 again. If we only perform the splitting phase, four segments would be resulted. However, note that the slopes of segment A_2 and segment B_1 are very similar. Hence, merging them is possible. After merging A_2 and B_1, three segments are remained. All of the segments now have different slopes. In other words, merging phase aims at combining all of the adjacent segments, provided that the mean square error, ε, would still be accepted by the t-test after merging. The hypothesis for the t-test is the same as Equation (2).

More formally, consider the time series T which has been transformed into another time series $T_{temp} = \{(t'_0, p'_0), (t'_1, p'_1), \ldots, (t'_m, p'_m)\}$ of length m after the splitting phase, such that $m \ll n$. Define $S_i = \{(t'_i, p'_i), (t'_{i+1}, p'_{i+1})\}$ as a segment in T_{temp}. If the null hypothesis over two adjacent segments, S_i and S_{i+1}, is accepted, then these two segments are regarded as a *candidate merging pair*. Let $\mathcal{L}_{merge}$ be a list containing all of these candidate merging pairs. One of the candidate merging pair resides in $\mathcal{L}_{merge}$ would be selected to merge if merging of it would be resulted in the minimum increase in the total error norm. The whole process is executed continuously until the t-test over all of the segments on the time series is rejected, i.e. $\mathcal{L}_{merge} = \phi$. Algorithm 3 illustrates

TABLE I
A 2×2 CONTINGENCY TABLE SUMMARIZED THE DISTRIBUTION OF FEATURE f_j IN THE DOCUMENT COLLECTION. THIS TABLE COULD BE MODELED BY A χ^2 DISTRIBUTION WITH ONE DEGREE OF FREEDOM.

	#documents have f_j	#documents do not have f_j
Segment $= S_k$	case 1	case 3
Segment $\neq S_k$	case 2	case 4

the whole procedure of merging phase.

C. Useful News Stories Selection

In reality, many news stories are valueless in prediction, i.e. they do not contribute to the prediction of the stock prices. In this section, we present how to select the valuable news stories.

Define $features$ to be any words in the news story collection. Let f_j be a feature in the news story collection. Recall that news story that are released within a segment are aligned back to that segment, i.e. $d_i \in \{\mathcal{D}_{S_k} \mid t_{rel}(d_i) \geq t_{begin}(S_k)$ and $t_{rel}(d_i) < t_{end}(S_k)\}$. By counting the presence or absence of f_j appearing during a given segment, a statistic model for discrete events could be formulated. In such model, the frequency of any feature appearing within the news story collection would be random with unknown distribution. In a model that features are emitted at a random process, two assumptions could be made: 1) The process of generating the features is *stationary*; and 2) The occurrence of every feature is *independent* of each other, i.e. $P(f_a) = P(f_a|f_b)$.

For the first assumption, if a feature is stationary, then in any arbitrary period, the probability of getting it is the same as at any other periods. In other words, if the probability of a feature appearing in some periods change dramatically, we can conclude that this feature exhibit an abnormal behavior in those periods, and it would be regarded as an important feature in there. Specifically, by counting the number of documents that: 1) contains feature f_j and is in Segment S_k; 2) contains feature f_j but is not in Segment S_k; 3) does not contain feature f_j but is in segment S_k; and 4) does not contain feature f_j and is not in segment S_k, a 2×2 contingency table could be formulated (Table I). Note that this table could be modeled by a χ^2 distribution with one degree of freedom.

For the second assumption, it is known as the independent assumption of feature distribution, which is a common assumption in text information management, especially for information retrieval, clustering and classification. Researches show that this assumption will not harm the system performance [30], [31], [32], [33]. Indeed, maintaining the dependency of features is not only extremely difficult, but also may easily degrade the system performance [34], [32], [33], [35].

For each feature f_j under each segment S_k, we calculate its χ^2 value, i.e. $\chi^2(f_j, S_k)$. If it is above a threshold, α, i.e. $\chi^2(f_j, S_k) \geq \alpha$, we conclude that the occurrence of feature f_j in segment S_k is significant, and this feature is appended into a feature list, $\mathcal{L}_{feature,S_k}$. $\mathcal{L}_{feature,S_k}$ stores all the features in which their occurrence in segment S_k are significant:

$$f_j \in \mathcal{L}_{feature,S_k} \text{ if } \chi^2(f_j, S_k) \geq \alpha \quad (5)$$

Define $\mathcal{D}_R$ and $\mathcal{D}_D$ be two sets containing the documents that support the rise movement and drop movement, respectively. Hence, these two sets are served as the positive training examples for the rise and drop trends. A document, d_i, which belongs to segment S_k ($d_i \in \mathcal{D}_{S_k}$), would be assigned to $\mathcal{D}_R$ if and only if the slope of S_k is positive and d_i contains a feature listed in $\mathcal{L}_{feature,S_k}$ (i.e. $d_i \in \mathcal{D}_R$ iff $f_j \in \{\mathcal{L}_{feature,S_k}$ and $d_i \in \mathcal{D}_{S_k}\}$). Similar strategy applies to $\mathcal{D}_D$.

Note that for $\chi^2 = 7.879$, there is only a probability of 0.005 that a wrong decision would be made such that a feature from a stationary process would be identified as not stationary, i.e. a random feature is wrongly identified as a significant feature. Hence, α is set to 7.879. Besides, only the features that appear in more than one-tenth of the documents in the corresponding period would calculate their χ^2 value. This is because rare features are difficult to estimate correctly and this can reduce significant computational cost. Algorithm 4 outlines the procedure of selecting positive training news stories.

Algorithm 4 select($\mathcal{D}$, $\mathcal{T}'_m$) – select positive training examples from a collection of documents $\mathcal{D}$ given a segmented time series $\mathcal{T}'_m$.

```
for each S_k in T'_m do
    L_{feature,k} = φ;
    D_{S_k} = φ;
    if t_rel(d_i) ≥ t_begin(S_k) and t_rel(d_i) < t_end(S_k) then
        Assign d_i to D_{S_k};
    end if
end for
for each S_k in T'_m do
    for each f_i ∈ D_{S_k} do
        if χ²(f_i) ≥ α then
            append f_i to L_{feature,k};
        end if
    end for
end for
D_R = φ;
D_D = φ;
for each f_j ∈ L_{feature,k} do
    if f_j ∈ {d_i | d_i ∈ D_{S_k}} then
        if slope(S_k) > 0 then
            Assign d_i to D_R;
        else
            Assign d_i to D_D;
        end if
    end if
end for
```

D. System Learning

Recall that in our training data, two types of data are available: $\mathcal{D}_R$ and $\mathcal{D}_D$. $\mathcal{D}_R$ and $\mathcal{D}_D$ represents the training documents correspond to the rise trend and the drop trend,6

respectively. Let N_R and N_D be the number of documents in $\mathcal{D}_R$ and $\mathcal{D}_D$, respectively. For the sake of simplicity, let us define $X \in \{R, D\}$.

Following the common practise of document preprocessing, for each document in X, $d_{j,X}$,a vector space model is constructed to represent it [33]:

$$d_{j,X} = \langle f_0 : w_{0,X}, f_1 : w_{1,X}, \ldots, f_n : w_{n,X} \rangle \quad (6)$$

where f_i is the i^{th} feature in $\mathcal{D}$ and $w_{i,X}$ is the weight of f_i in X. $w_{i,X}$ indicates the importance of f_i in $d_{j,X}$. Follow the existing works, we use a $tf \cdot idf$ schema for calculating the weights [36], [33]:

$$w_{i,X} = \begin{cases} tf_{i,j} \cdot \log_{N_X} \dfrac{N_X}{df_{i,X}} & \text{if } df_{i,X} \neq 0, \\ 0 & \text{if } df_{i,X} = 0. \end{cases} \quad (7)$$

where $tf_{i,j}$ is the *term frequency* (i.e. the number of times f_i appears in d_j) and $df_{i,X}$ is the *document frequency* (i.e. the number of documents contains f_i in X). Finally, $w_{i,X}$ is normalized to unit length so as to account for the differences in the length of each document.

In this formulation, each feature is regarded as a single dimension and the weight of the feature is regarded as the coordinate for that dimension. In other words, each document has n-dimension ($\Re^n$), where n is the total number of features in $\mathcal{D}$. Thus, our training data consists N_X pairs of $(d_{1,X}, y_{1,X}), (d_{2,X}, y_{2,X}), \ldots, (d_{N_X,X}, y_{N_X,X})$, with $d_{i,X} \in \Re^n$ and $y_{i,X} \in \{-1, 1\}$. This problem then reduced to a two class pattern recognition problem in which we are trying to find a hyperplane:

$$f : \mathcal{D}_X^T \beta_X + \beta_{0,X} = 0 \quad \text{and} \quad \| \beta_X \| = 1 \quad (8)$$

which maximize the *margin*, C_X, between the positive training examples in X and negative training examples in X. Thus, this problem reduced to the following optimization problem:

$$\max_{\beta_X, \beta_{0,X}, \|\beta_X\|=1} : \quad C_X \quad (9)$$

$$\text{subject to :} \quad y_{i,X}(d_{i,X}^T \beta_X + \beta_{0,X}) \geq C_X, \forall i \in X \quad (10)$$

By dropping the norm constraint on β_X, solving this problem is equivalent to solve the following optimization problem:

$$\min_{\beta_X, \beta_{0,X}} : \quad \| \beta_X \| \quad (11)$$

$$\text{subject to :} \quad y_{i,X}(d_{i,X}^T \beta_X + \beta_{0,X}) \geq 1, \forall i \in X \quad (12)$$

Since document vectors are very sparse, the two classes will share a high overlapping region in the feature space. To deal with it, slack variables, $\xi_X = (\xi_{1,X}, \xi_{2,X}, \ldots, \xi_{N_X,X})$, is introduced (Figure 5):

$$\min_{\beta_X, \beta_{0,X}} : \quad \frac{1}{2} \| \beta \|^2 + C \sum_{i=1}^{N_X} \xi_{i,X} \quad (13)$$

$$\text{subject to :} \quad y_{i,X}(d_{i,X}^T \beta_X + \beta_{0,X}) \geq 1 - \xi_{i,X}, \forall i \in X \quad (14)$$

$$\xi_{i,X} = 0, \forall i \in X \quad (15)$$

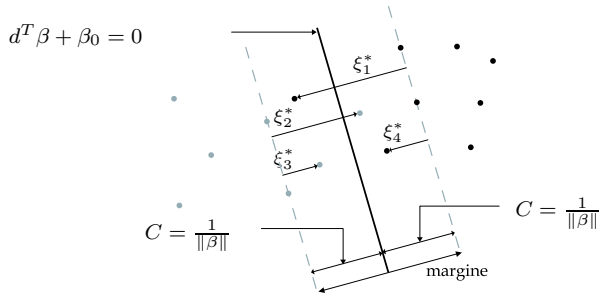


Fig. 5. The basic concept of the classifi er in a two-dimensional situation. The decision boundary is the solid line, while the broken lines bound the maximal margin of width $2C$. The points labeled ξ_j^* are on the wrong side of their margin by an amount $\xi_j^* = C\xi_j$. Points on the correct side have $\xi_j^* = 0$.

Constraint (14) requires that all training examples are classified correctly up to some slack $\xi_{i,X}$. If a training example lies on the wrong side of the hyperplane, the corresponding $\xi_{i,X}$ is ≥ 1. Therefore $\sum_{i=1}^{N_X} \xi_{i,X}$ is an upper bound on the number of training errors.

Consequently, solving this optimization problem is equivalent to solve a Support Vectors Machine (SVM) problem [37]. For computational reason, it is far more efficient to convert the above primal optimization problem to the Lagrangian (Wolfe) dual optimization problem [17], [37]:

$$\min : \quad L_X(\alpha) = -\sum_{i=1}^{N_X} \alpha_i + \frac{1}{2} \sum_{i=1}^{N_X} \sum_{j=1}^{N_X} Q_{ij} \quad (16)$$

$$\text{subject to :} \quad \sum_{i=1}^{N_X} \alpha_{i,X} y_{i,X} = 0 \quad (17)$$

$$Q_{ij} = \alpha_{i,X} \alpha_{j,X} y_{i,X} y_{j,X} x_{i,X}^T x_{j,X} \quad (18)$$

$$0 \leq \alpha_{i,X} \leq C, \forall i \in X \quad (19)$$

The size of the optimization problem depends on the number of training examples N. Defining a matrix $Q = y_i y_j x_i^T x_j$. Note that the size of Q is around N^2. For learning task with thousand of features and thousand of training documents, it becomes impossible to keep Q in memory. Standard implementations require either explicit storage of Q or re-compute Q every time when it is needed. However, this becomes prohibitively expensive. In this paper, we applied the technique describe by Joachims [38], which decompose the learning task into several sub-tasks. The solution of β_X, $\beta_{0,X}$ and $\xi_{i,X}$ can be computed as:

$$\beta_X = \sum_{i=1}^{N_X} \alpha_{i,X} y_{i,X} d_{i,X} \quad (20)$$

$$\beta_{0,X} = y_{sv,X} - \beta_X d_{sv,X} \quad (21)$$

$$\xi_X = \max\{1 - y_{i,X}(\beta_X d_{i,X} + \beta_{0,X}), 0\} \quad (22)$$

where the pair $(d_{sv,X}, y_{sv,X})$ must be a $support\ vector$ with $\alpha_{sv} < C$. Support vectors are those observations with the coefficient $\alpha_{i,X} \neq 0$.

TABLE II
THE CATEGORIES POWERED BY REUTERS.

Category	Category	Category
Hotel	Industry (A-G)	Consultant (A-G)
Financial	Industry (H-O)	Consultant (H-O)
Properties	Industry (P-Z)	Consultant (P-Z)
Utility	Germany	Miscellaneous

E. System Operation

After the system training process, two classification models are generated in which they are responsible for determining whether an unseen document would trigger the rise event and drop event, respectively. Given the solutions of β_X and $\beta_{0,X}$, the decision function for any unseen document, $\hat{d}$, can be written as:

$$\begin{aligned} G_X(\hat{d}) &= \text{sign}[f(\hat{d})] \\ &= \text{sign}[\hat{d}^T\beta_X + \beta_{0,X}] \end{aligned} \tag{23}$$

If $G_R > 0$ ($G_D < 0$), then the unseen document, $\hat{d}$ is classified as triggering the rise event (drop event). In other words, $\hat{d}$ is believed to affect the time series such that it goes upward (downward). If both G_R and G_D are < 0, then $\hat{d}$ is classified as noise which means that $\hat{d}$ does nothing with the time series. If both G_R and G_D are > 0, then the actual impacts of $\hat{d}$ is ambiguous since it triggers both the rise and drop events, which is impossible. In such a case, we would ignore $\hat{d}$ also, and classified it as noise as well.

IV. EVALUATION

A prototype system using Java™ is developed to evaluate the proposed system. All of the experiments are conducted on a Sun Blade-1000 workstation running Solaris 2.8 with 512MB physical memory and with a 750MHz Ultra-SPARC-III CPU. Intra-day stock prices and real-time news stories are archived through Reuters Market 3000 Extra[7] from 20^{th} January 2003 to 20^{th} June 2003. All data are stored into IBM DB2 Version 7.1[8].

For the real-time news stories, there are more than 350,000 documents archived. Note that Reuters has assigned to which sectors, countries, etc, the news stories should belong. Therefore, we do not need to worry about how these news stories should be organized. All features from the news stories are stemmed and converted to lower cases, in which punctuation and stop-words are removed, numbers, web page addresses and email addresses are ignored.

For the stock data, intra-day stock prices of all the Hong Kong stocks are recorded[9]. The stocks belong to one of the categories listed in Table II. According to the observations given by the technical analysis that price movements associated with light volumes denotes only temporal movements, but not trends, thus, for each stock, the transactions that are associated with light volumes (e.g. few hundred shares) are ignored. In order to account for the different price range of different stocks, stock prices of all stocks are normalized.

A. Time Series Evaluations

Figure IV-A shows the typical results after applying the t-test based split and merge segmentation algorithm on three stocks. Due to the space limited, we report only three cases. Other stocks behave in the similar way. The reported stocks are: 1) Cheung Kong (0001.HK); 2) Cathay Pacific (0293.HK) and 3) TVB (0511.HK). The unmodified stock data are shown on the top while the segmented data are shown on the bottom. We can see that the trends generated are quite reasonable and suitable. Note that the longest trend lasts for 2 weeks while the shortest one lasts for 3 days. This means that all of the trends generated are tertiary movements.

V. PREDICTION EVALUATIONS

One of the best way to evaluate the reliability of a prediction system is to conduct a market simulation which mimics the behaviors of investors using real-life data. As a result, two market simulations are conducted:[10]

- **Simulation 1: Proposed System**: Shares are bought or sold based solely on the content of the news stories. Two strategies are adopted:
  - For each stock, if the prediction of its upcoming trend is positive, then shares of it are bought immediately. The shares would be sold after holding for m day(s).
  - For each stock, if the prediction of its upcoming trend is negative, then shares of that stock are sold for short. The shares would be bought back after m day(s).

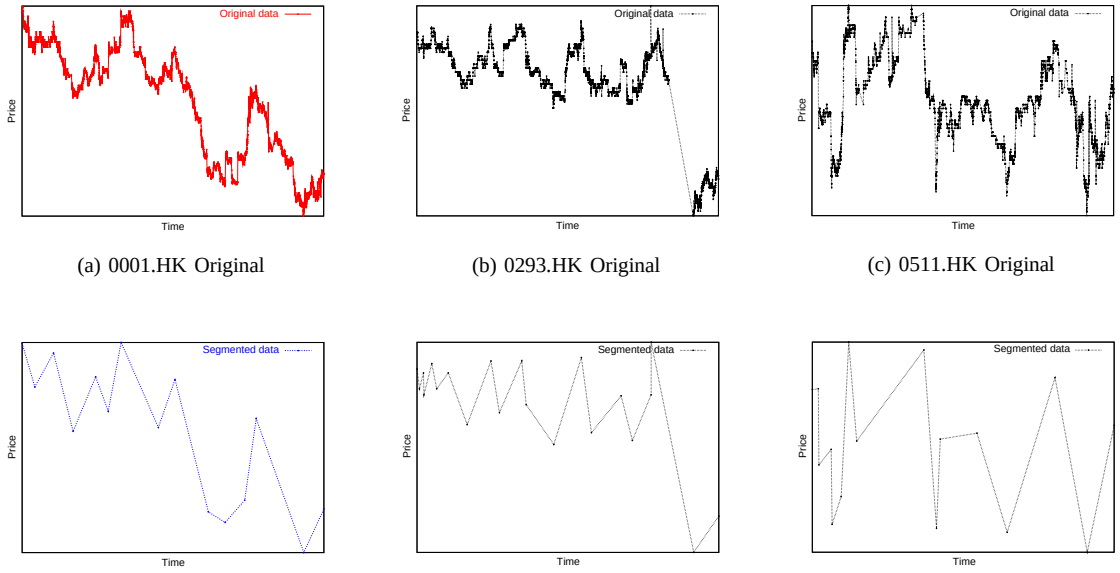
  An analysis of how m affects the evaluation results is given in Section V-B. In this section, m is set to 3 working days for simplicity. If the market is closed when the decision is made, then shares will be bought or sold in the beginning of the next active trading day.
- **Simulation 2: Buy-and-Hold Test**: For each stock, shares of that stock is bought at the beginning of the evaluation period. At the end of the evaluation period, all of the shares remain on hand are sold. This simulation serves as a base-line comparison which is used to demonstrate the *do-nothing strategy*.

In the above market simulations, rate of return, r, is calculated. As a result, how much shares are bought in each transaction could be ignored.

A. Simulation Results

Table III shows the results of the market simulations. From the table, Simulation 1 far outperforms Simulation 2. In order to see whether the earnings from the proposed system are statistically significant, another 1,000 simulations are conducted. In these simulations, the decisions of buying and selling were made at the same time as the proposed

[7] http://www.reuters.com
[8] http://www.ibm.com
[9] The stocks which have too few transaction records are ignored. This is simply because there are not enough data for training and/or evaluation.
[10] The assumption of zero transaction cost is carried out



(a) 0001.HK Original (b) 0293.HK Original (c) 0511.HK Original

(d) 0001.HK Segmented (e) 0293.HK Segmented (f) 0511.HK Segmented

Fig. 6. Before and after applying the t-test based split and merge segmentation algorithm. On the top: The original time series. On the bottom: The segmented time series. Three stocks are selected to report here: (1) Cheung Kong (0001.HK); (2) Cathay Pacific (0293.HK); and (3) TVB (0511.HK).

TABLE III

THE OVERALL EVALUATION RESULTS OF THE TWO MARKET SIMULATION. HERE, r IS THE RATE OF RETURN.

	Simulation 1	**Simulation 2**
Accumulative r	**18.06**	-20.56
Stand. Dev. of r	3.40	**2.15**
Maximum r	**12.42**	2.21
Minimum r	**-9.83**	-18.10
Top ten average r	**8.18**	1.11
Least ten average r	**-3.69**	-18.56

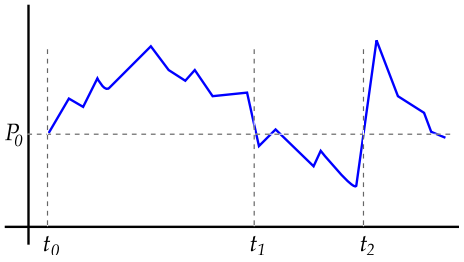


Fig. 7. A simple diagram illustrates the meaning of hit rate.

system, but without referencing to the contents of the news stories, i.e. the decisions are random. We then compare the cumulative earnings that are produced from the randomized trials and Simulation 1. For the randomized system, there are only 78 out of 1,000 trials that have rate of return exceed our proposed work. Thus, the proposed system is significant at the 0.5% level.

B. Hit Rate Analysis

Hit rate is another important measurement for the predictability of a forecasting system, especially for those kind of systems that are similar to our proposed one. Hit rate analysis can indicate how often the sign of return is correctly predicted. Figure 7 illustrates this idea. Assume that at t_0 a prediction which states that the stock price will go upward is made. Since from t_0 to t_1 (T_1), the stock prices are above p_0, we conclude that the prediction is correct in this period, i.e. *hit*. However, from t_1 to t_2 (T_2), the stock prices are below p_0, we therefore conclude that the prediction is wrong in T_2, i.e. *missed*. Thus, if the prediction period is varied, different conclusion could be drawn. In other words, the value of m in the market simulation presented in Section V-A is a critical factor.

TABLE IV

THE HIT RATE OF THE PROPOSED SYSTEM BY VARYING HOLDING PERIOD. HERE, THE RETURN IS CALCULATED IN RATE OF RETURN.

	Hit Rate	**Acc. Return**	**S.D. of Return**
1 day ($m = 1$)	51.0%	6.58	**1.147**
3 day ($m = 3$)	61.6%	18.06	3.400
5 day ($m = 5$)	**65.4%**	**21.49**	4.135
7 day ($m = 4$)	55.7%	7.22	3.791

Table IV shows the hit rate and the rate of return of the proposed system by varying the value of m. The accumulative return and hit rate both increase as m increase. It suggests that the system is most stable and suitable for applying the prediction within 3-5 days. It also suggests that such kinds of movements should be tertiary movements.

A careful examination of the prediction results would realize

that one of the major reasons for making error is that two news stories may be very similar in their contents, but have totally different implications.

VI. Conclusion

Scholars and professionals from different areas have shown that there is a high relationship between the news stories and the behaviors of the financial markets. In this paper, we revisit the problem and use real-time news stories and intra-day stock prices for our study. These data are chosen because they are readily available and the evaluation results obtained can easily be verified.

Several data mining and text mining techniques are incorporate in the system architecture. The tertiary movements on the stock price movements are identified by a novel piecewise linear approximation approach: a t-test based split and merge segmentation algorithm. News stories are aligned to the stock trends basic on the idea of Efficient Market Hypothesis. A document selection heuristics that is based on the χ^2 estimation is used for selecting the positive training documents. Finally, the relationship between the contents of the news stories and trends on the stock prices are learned through support vectors machine. Different experiments are conducted to evaluate various aspect of the proposed system. In particular, a market simulation using real-life data is conducted. Encouraging results are obtained in all of the experiments. Our study show that there is a high relationship between news stories and the movements of stock prices. Furthermore, by monitoring this relationship, actionable decisions could be made also.

References

[1] P. A. Adler and P. Adler, "The market as collective behavior," in *The Social Dynamics of Financial Markets*, P. A. Adler and P. Adler, Eds. Jai Press Inc., 1984, pp. 85–105.

[2] H. Blumer, "Outline of collective behavior," in *Readings in Collective Behavior*, 2nd ed., R. R. Evans, Ed. Chicago: Rand McNally College Pub. Co, 1975, pp. 22–45.

[3] J. M. Clark, "Economics and modern psychology," *Journal of Political Economy*, vol. 26, pp. 136–166, 1918.

[4] L. Tvede, *The Psychology of Finance*, revised ed. John Wiley and Sons, Inc., 2002.

[5] L. Festinger, *A theory of cognitive dissonance*. Stanford, Calif.: Stanford Univesity Press, Reprinted in 1968.

[6] M. Klausner, "Sociological theory and the bechavio of financial markets," in *The Social Dynamics of Financial Markets*, P. A. Adler and P. Adler, Eds. Jai Press Inc., 1984, pp. 57–81.

[7] G. P. C. Fung, J. X. Yu, and W. Lam, "News sensitive stock trend prediction," in *Proceedings of the 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Taipei, Taiwan, 2002, pp. 289–296.

[8] ——, "Stock prediction: Integrating text mining approcah using real-time news," in *Proceedings of the 7th IEEE International Conference on Computational Intelligence for Financial Engineering*, Hong Kong, China, 2003, pp. 395–402.

[9] V. Lavrenko, M. D. Schmill, D. Lawire, P. Ogivie, D. Jensen, and J. Allan, "Mining of concurrent text and time series," in *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining Workshop on Text Mining*, Boston, MA, USA, 2000, pp. 37–44.

[10] T. Fawcett and F. J. Provost, "Activity monitoring: Noticing interesting changes in behavior," in *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining*, San Diego, California, USA, 1999, pp. 53–62.

[11] D. Permunetilleke and R. K. Wong, "Currency exchange rate forecasting from news headlines," in *Proceedings of the 13h Australian Database Conference*, Melbourne, Australia, 2002, pp. 131–139.

[12] J. D. Thomas and K. Sycara, "Integrating genetic algorithms and text learning for financial prediction," in *Proceedings of the Genetic and Evolutionary Computing 2000 Conference Workshop on Data Mining with Evolutionary Algorithms*, Las Vegas, Nevada, USA, 2000, pp. 72–75.

[13] B. Wuthrich, D. Permunetilleke, S. Leung, V. Cho, J. Zhang, and W. Lam, "Daily prediction of major stock indices from textual www data," in *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, New Youk, USA, 1998, pp. 364–368.

[14] R. J. Bauerm Jr and J. R. Dahlquist, *Technical Market Indicators*. John Wiley and Sons, Inc., 1999.

[15] R. D. Edwards and J. Magee Jr, *Technical Analysis of Stock Trends*, 5th ed. Springfield, 1966.

[16] S. A. Nelson, *The ABC of Stock Market Speculation*, 3rd ed. Fraser Publishing, 1903.

[17] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2002.

[18] F. Klein and J. A. Prestbo, *News and the Market*. Chicago: Henry Regenry, 1974.

[19] K. Nigam, J. Lafferty, and A. McCallum, "Using maximunm entropy for text classification," in *Proceeding of the 16th International Joint Conference Workshop on Machine Learning for Information Filtering*, Stockholm, Sweden, 1999, pp. 61–67.

[20] B. Wuthrich, "Probabilistic knowledge bases," *IEEE Transactions of Knowledge and Data Engineering*, vol. 7(5), pp. 691–698, 1995.

[21] ——, "Probabilistic knowledge bases," *International Journal of Intelligent Systems in Accounting Finance and Management*, vol. 6, pp. 269–277, 1997.

[22] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proceedings of the 21th International Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 1998.

[23] P. A. Adler and P. Adler, *The Social Dynamics of Financial Markets*. Jai Press Inc, 1984.

[24] W. J. Eiteman, C. A. Dice, and D. K. Eiteman, *The Stock Market*, forth ed. McDGraw-Hill Book Company, 1966.

[25] Y. Qu, C. Wang, and X. S. Wang, "Supporting fast search in time series for movement patterns in multiples scales," in *Proceedings of the 7th International Conference on Information and Knowledge Management*, Bethesda, Maryland, USA, 1998, pp. 251–258.

[26] C. Wang and X. S. Wang, "Supporting content-based searches on time series via approximation," in *Proceedings of the 12th International Conference on Scientific and Statistical Database Management*, Berlin, Germany, 2000, pp. 69–81.

[27] E. J. Keogh, S. Chu, D. Hart, and M. J. Pazzani, "An online algorithm for segmenting time series," in *Proceedings of the 1st IEEE International Conference on Data Mining*, San Jose, California, USA, 2001, pp. 289–296.

[28] T. Pavlidis and S. L. Horowitz, "Segmentation of plane curves," *IEEE Transactions on Computers*, vol. c23(8), pp. 860–870, 1974.

[29] R. O. Duda and P. E. Harts, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[30] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine Learning*, vol. 29(2-3), pp. 103–130, 1997.

[31] D. D. Lewis, "A sequential algorithm for training text classifiers," in *Proceedings of the 17th International Conference on Research and Development in Information Retrieval*, Dublin, Ireland, 1994, pp. 3–12.

[32] ——, "The independence assumnption in information retrieval," in *Proceedings of the 10th European Conference on Machine Learning*, Chemnitz, Germany, 1998, pp. 4–15.

[33] F. Seabastiani, "Machine learning in automated text categorization," *ACM Computing Surverys*, vol. 34(1), pp. 1–47, 2002.

[34] W. B. Croft, "Boolean queries and term dependencies in probabilistic retrieval models," *Journal of the American Society for Information Science*, vol. 37(2), pp. 71–77, 1983.

[35] C. J. van Rijsbergen, "A theoretical basis for the use of co-occurrence data in information retrieval," *Journal of Documentation*, vol. 33(2), pp. 106–119, 1977.

[36] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Process Management*, vol. 24(5), pp. 513–523, 1998.

[37] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995.

[38] T. Joachims, "Making large-scale svm learning practical," Computer Science Department, University of Dortmund, Tech. Rep. LS-8 (24), 1998.

An On-Line Web Visualization System with Filtering and Clustering Graph Layout

Wei Lai, Xiaodi Huang, Ronald Wibowo, and Jiro Tanaka

***Abstract*—A Web graph refers to the graph that is used to represent relationships between Web pages in cyberspace, where a node represents a URL and an edge indicates a link between two URLs. A Web graph is a very huge graph as growing with cyberspace. To use it for Web navigation, only a small part of the Web graph is displayed each time according to a user's navigation focus. The graph layout has always been a challenge for visualizing systems. In this paper, we present a visualization system of an online Web graph, together with the methods for clustering and filtering large graphs. In this system, a Web crawler process is used to get on-line information of the Web graph. Filtering and clustering processes reduce the graph complexities on visualization. In particular, the filtering removes those unimportant nodes while the clustering groups a set of highly connected nodes and edges into an abstract node. The visualization process incorporates graph drawing algorithms, layout adjustment methods, as well as filtering and clustering methods in order to decide which part of the Web graph should be displayed and how to display it based on the user's focus in navigation.**

***Index Terms*— Graph visualization, Filtering, Clustering, Web graph**

I. Introduction

The amount of information now available through the World Wide Web (WWW) has grown explosively. An increasing number of tools are available to assist users to manage and access information on the WWW, such as Netscape and Internet Explorer. The key requirement for a Web browser is to show the details for the users' focused information and to facilitate navigation within the whole information hyperspace. It is, however, impossible to display this huge and growing hyperspace for users to get its whole structure in helping navigation. The navigation approach used in most Web browsers is simply from one page to another page. Although current Web browsers can provide bookmarks and history lists in a linear way, they cannot show relationships between the URLs.

Some researchers have proposed "site mapping" methods [3, 12, 15] in an attempt to find an effective way of constructing the structured geometrical map for a Web site (i.e. a local map). However, this map can only guide users through a very limited region of cyberspace, and does not help the users in their overall journey through the cyberspace.

Other attempts use a graph for the WWW navigation. The whole cyberspace of the WWW is regarded as a Web graph [6, 9, 10]. In the Web graph, a node represents a Web page's URL and an edge represents a link between two URLs. This approach is placed an emphasis on navigation, but ignores achievement of a better local view for the site mapping. The graph layout by this approach shows all possible hyperlinks and makes the layout look so messy. This makes a site-mapping view sometimes unclear to users.

The primary difficulty for creating an auto-generated sitemap lies in that the number of the links can be quite big, or even huge. The presentation of these links will become messy and hard to read, so that the visualization will become useless.

This paper presents an on-line Web visualization system by using filtering and clustering to reduce visual complexities of the Web graph. The system includes the processes of Web crawler, filtering and clustering, and visualization. The Web crawler process is used to get on-line information of the Web graph. Filtering and clustering processes reduce the graph complexities on visualization. The filtering is used to remove those unimportant nodes, and the clustering is used to make a set of nodes and edges (a sub-graph) to an abstract node. The visualization process uses graph drawing algorithms and layout adjustment algorithms for graph layout. We begin with the description of our system in the following section, and then present the filtering and clustering methods in Sections 3 and 4. A case study is provided in Section 5, followed by the conclusion in Section 6.

II. The On-line Web visualization system

The on-line Web visualization system with filtering and clustering graph layout (we call it the FCG system) supports a user to use a graph to navigate the cyberspace. The Web graph is a very huge graph as the cyberspace keeps growing. During the Web navigation, each time only a small part of the Web graph is displayed. We call it a sub-Web graph which is formed based on the user's focus in navigation.

Figure 1 shows the FCG system in action. According to the user's choice of a node in navigation, the relevant Web page is shown up.

The user can navigate the Web graph by selecting a node. This selected node is called the focused node. The system can smoothly add some new nodes which are closed to the focused node and remove some other nodes which are far away from the

Manuscript received September 15 , 2004.

Wei Lai and Ronald Wibowo are with School of Information Technology, Swinburne University of Technology, PO Box 218, Hawthorn, VIC 3122, Australia (wlai@swin.edu.au).

Xiaodi Huang is with Department of Mathematics and Computing, The University of Southern Queensland, Australia.

Jiro Tanaka is with Institute of Information Sciences and Electronics, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan.

focused node with the filtering and clustering processes based on the size of a display window.

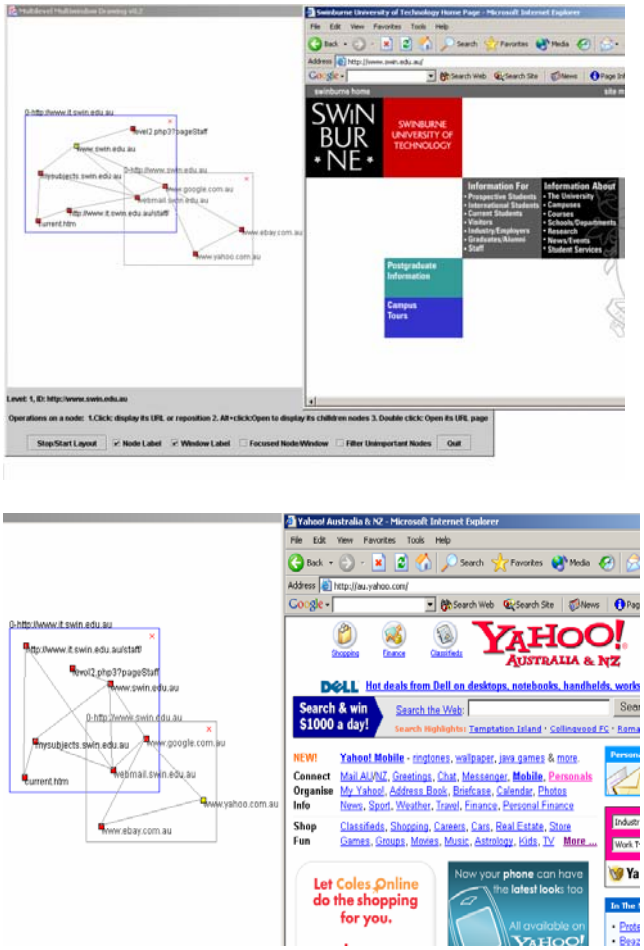


Fig. 1. Design of the FCG System

The design of the FCG system is as follows. A Web crawler extracts on-line URLs relationships from the cyberspace and constructs the Web graph represented in a text file format. This file is processed by the filtering and clustering and then goes to the visualization process for graph layout. The user can interact with the system to adjust the filtering, clustering, and visualization.

The FCG system has three modules. Each module is treated differently and can be implemented individually. The first module, called the Web crawler, is to obtain the hyperlinks among Web sites as mentioned above. The crawler crawls from a given input URL and stop when the defined depth is reached. The Web crawler then saves the URLs list into a text file. The second module is about the filtering/clustering process, while the third module includes the visualization process. The overall process is detailed in Figure 2.

In this system, we integrate the techniques of the Web crawler [2, 4, 13], graph drawing algorithms [1], layout adjustment methods [14], and the filtering and clustering methods.

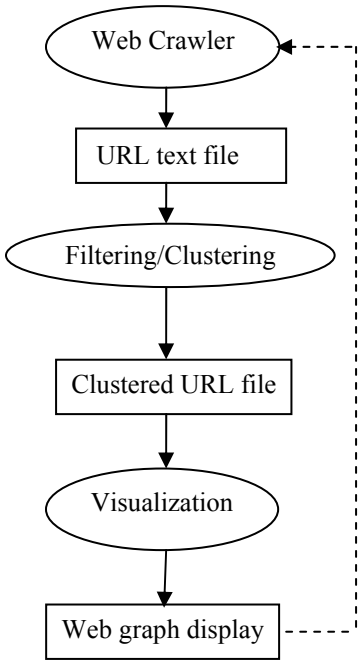


Fig. 2. Design of the FCG System

Each module can run independently with the given input, and it also produces an output. The dashed line in Figure 2 implies that the Web graph is changed on the basis of the results of the Web crawler. In other words, the new Web pages collected by the Web crawler can immediately be reflected in the updated Web graph.

As mentioned before, the Web crawler in Figure 2 is employed to extract the links from a given URL Web site, with a specified depth of exploration. The detailed process of this Web crawler can be illustrated in Figure 3.

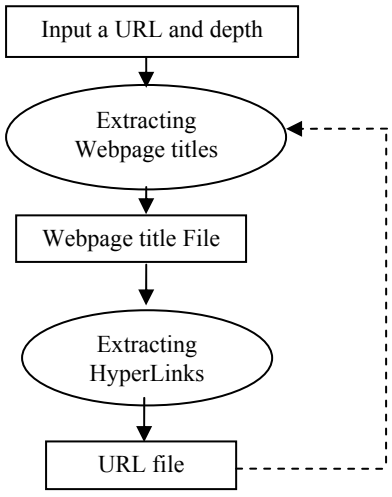


Fig.3. Web crawler process

Note that a dashed line in Figure 3 between “URL file” and

"Extracting Webpage titles" means that the process will continue until reaching the given exploration depth. For example, if the given depth is one then the process will only run once. If the given depth is three, then the process will be carried out three times with the immediately previously crawled URLs as new starting points of the exploration.

III. Filtering

To enhance the readability of the layout, some filter mechanisms are applied to the Web graph. The filter is to reduce the size of the graph by removing weak links, defined as those edges with connected nodes whose degrees are less than a predefined number. The use of the filter makes the Web graph layout easier to read due to reduction of the number of nodes and edges.

Usually, a large graph is automatically generated from an information source. This may unavoidably lead to the creation of "noise" information. For example, the use of a Web crawler program easily extracts some unwanted image files, together with html Web pages, from a Web site when constructing a Web graph. In addition, Filtering can suppress unimportant nodes and their related edges to highlight those important nodes by using an adjustable threshold to control appearances of the nodes.

The purpose of the FCG system is to display a Web graph for users to explore. In the following example of a graph in Figure 5 (this graph is a sub-Web graph), the links to CSS files and to image files are therefore considered to be unimportant links. Figure 6 shows an example by applying the filtering to remove these links.

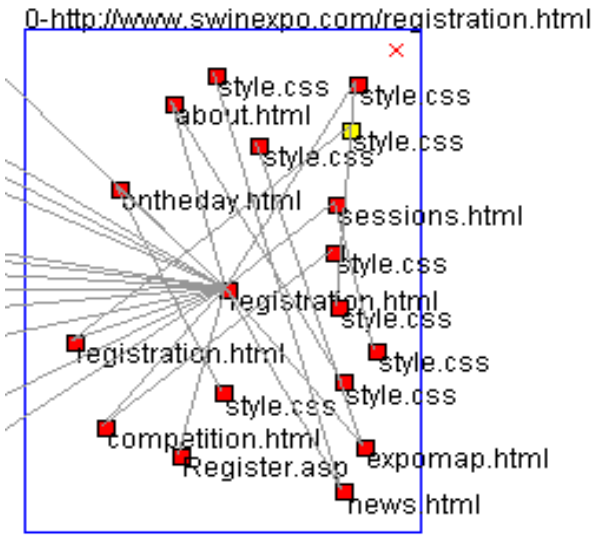


Fig.5. A graph obtained by the Web crawler process

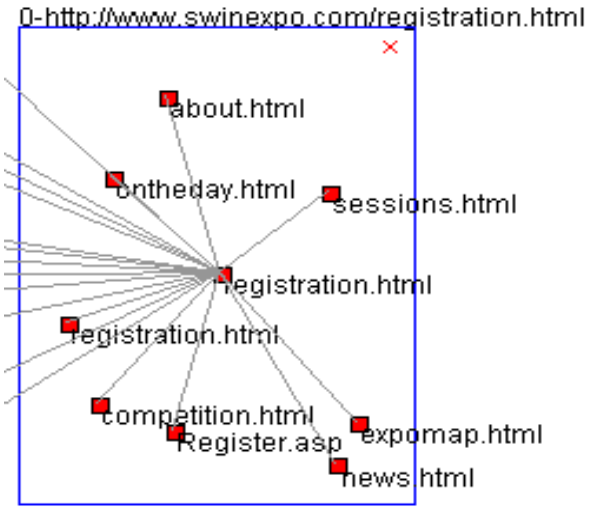


Fig.6. The graph obtained after the filtering

A filtering algorithm [8] is applied directly the Web graph. This algorithm is to calculate the node rank values for every nodes based on their connection degrees with other nodes, geodesic distances with other nodes, and the "intermediary" important role between the various other nodes. The range of the node rank value is between 0 and 1. In practice, with an appropriate threshold, some "noise" nodes or less important nodes are removed.

IV. Clustering

In the implementation, the filtering and the clustering as a whole module starts by accepting an input text file, produced by the Web crawler, and ends with outputting a file containing a list of clustered URLs. The clustering procedure is shown in Figure 7.

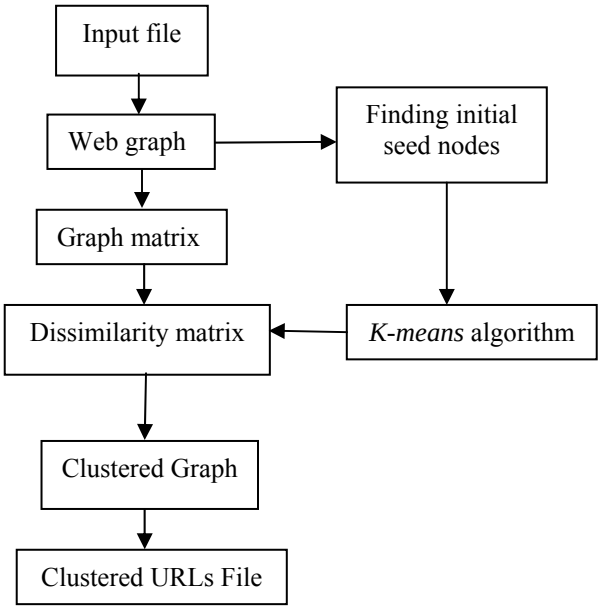


Fig. 7. Clustering process

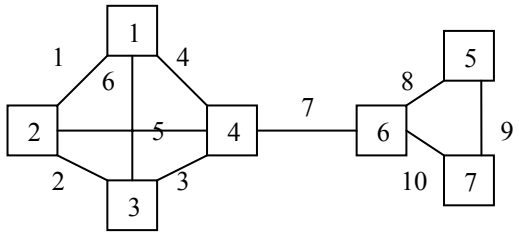
At the start of the process, a Web graph is constructed from the URL file. This graph is then represented as an edge-by-node matrix, where each column indicates a node encoded by its connecting edges, and each row represents an edge characterized by its related nodes. An example is shown in Figure 8. Note that after filtering, the graph is still represented by this kind of matrix.

The purpose of clustering a graph is to find relatively highly connected sub-graphs in a graph. With the matrix as the representation of the graph, a similarity matrix can be derived as to measure the degree of connection between two nodes.

The algorithm for clustering the elements in the matrix [7] for a given graph G can be summarized briefly as follows:

Input: G (V, E)
Output: A clustered graph

- Construct the edge-by-node incident matrix R.
- Calculate similarity values of neighbour nodes pair according to the formula (1).
- Find the shortest paths between non-neighbour node pairs by using the Dijkstra's algorithm.
- Construct the node similarity matrix according to the formula (2)
- Apply the k-means algorithm to the dissimilarity matrix constructed by the formula (3)



$$Node\ \begin{matrix}1&2&3&4&5&6&7\end{matrix}$$

$$R=\begin{bmatrix}1&1&0&0&0&0&0\\0&1&1&0&0&0&0\\0&0&1&1&0&0&0\\1&0&0&1&0&0&0\\0&1&0&1&0&0&0\\1&0&1&0&0&0&0\\0&0&0&1&0&1&0\\0&0&0&0&1&1&0\\0&0&0&0&1&0&1\\0&0&0&0&0&1&1\end{bmatrix}\begin{matrix}1\\2\\3\\4\\5\\6\\7\\8\\9\\10\end{matrix}\quad Edge$$

Fig. 8. An example of a graph and its edge by node matrix

$$sim(r_i, r_j) = \frac{r_i^T r_j}{r_i^T r_i + r_j^T r_j - r_i^T r_j}$$

$$sim(r_i, r_j) = \frac{(R^T e_i)(e_j^T R)}{(R^T e_i)(e_i^T R) + (R^T e_j)(e_j^T R) - (R^T e_i)(e_j^T R)} \quad (1)$$

where i and j = 1, …, n, and e_i (or e_j) denotes the i-th (or j-th) canonical vector of dimension e, i.e., $e = (1,1,\ldots,1)^T$.

This expression is used to calculate the similarity values between two neighbour nodes. In the case of non-neighbour node, the following formula is employed instead.

$$sim(r_i, r_j) = \begin{cases} \max_{P'} \{ \prod_{(r_m, r_n) \in P} sim(r_m, r_n) \} & if\ P \neq \phi \\ 0 & if\ P = \phi \end{cases} \quad (2)$$

where P is a set of pairs of nodes in the shortest path between nodes i and j, namely $P = \{(i,m),\cdots,(k,j)\}$. Such several possible shortest paths consist of a set P'

For example, the symmetric similarity matrix of Figure 8 is shown in Figure 9. The formula (2) is in fact a metric of measure the degree of two nodes in a graph, which satisfies fundamental properties: non-negativity, symmetry and ranging within [0, 1].

$$S=\begin{bmatrix}1.000&&&&&&\\0.200&1.000&&&&&\\0.200&0.200&1.000&&&&\\0.167&0.167&0.167&1.000&&&\\0.007&0.007&0.007&0.042&1.000&&\\0.028&0.028&0.028&0.167&0.253&1.000&\\0.007&0.007&0.007&0.042&0.333&0.250&1.000\end{bmatrix}$$

Figure 9: The node similarity matrix of the graph in Figure 8.

From similarity matrix S, we can easily obtain the dissimilarity matrix D by the following formula (3), which will be used as an input of the k-means algorithm.

$$D = [1]_{n \times n} - S \quad (3)$$

The clustering process produces a text file containing many lines, with each line describing the links between two URLs, and between two clustered groups.

If we apply the clustering process to a graph several times, the multi-level abstraction of clustered graphs will be generated. Figure 10 shows the result of applying the clustering to the graph in Figure 8.

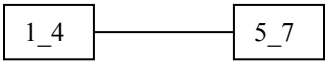


Fig. 10. An example of applying the clustering

The time requirements for calculation of the similarity matrix is O($e + n \log n$) while the running time for the k-means is O(rkn), where e is the number of edges, n is the number of nodes r is the number of iteration, and k is the number of desired cluster. Thus the total time complexities will be O ($e + n \log n + rkn$).

V. A CASE STUDY

For the purposes of this case study, we restrict our interest to visually navigate http://www.swin.edu.au (Swinburne university Website) with two levels of exploration depth using the FCG system. Specifically, our goal is to evaluate the FCG system in producing visualizations that can be used properly. We investigate the drawing, representation of a Web graph produced by the FCG system.

That is, we focus on three issues:

- The visualization of Swinburne university Website as a Web graph.
- Discussion of the drawings and representations, in terms of their ability to navigate the Web graph.
- The measurement of performance of the FCG.

A. The Experiments

To investigate the FCG system in producing Web graph visualization, we tested the FCG system to view Swinburne Website with two levels of exploration depth.

The results of this case study are presented in the following two sections. First, we present a picture gallery of different layouts produced by the FCG system. Second, we present the discussion and performance measures of the layouts shown in the picture gallery.

B. Picture Gallery

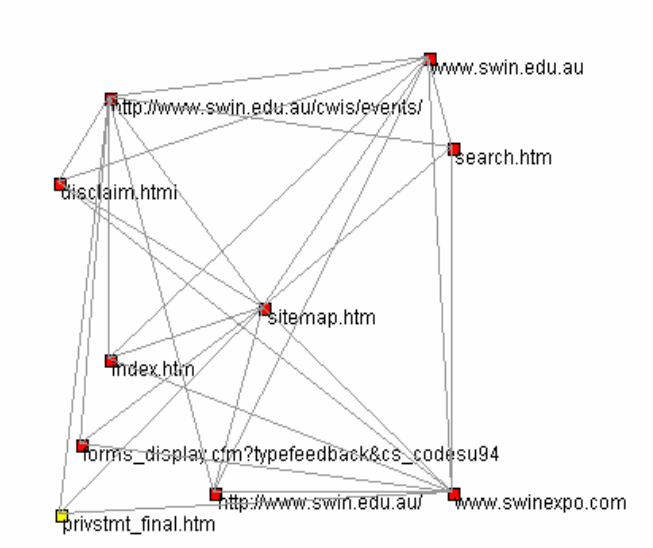


Fig.11. The Swinburne Web site using FCG

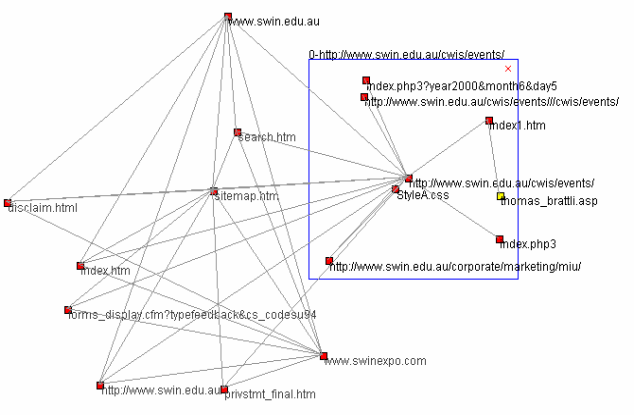


Fig.12. Expanded node labeled as http://www.swin.edu.au/cwis/events/

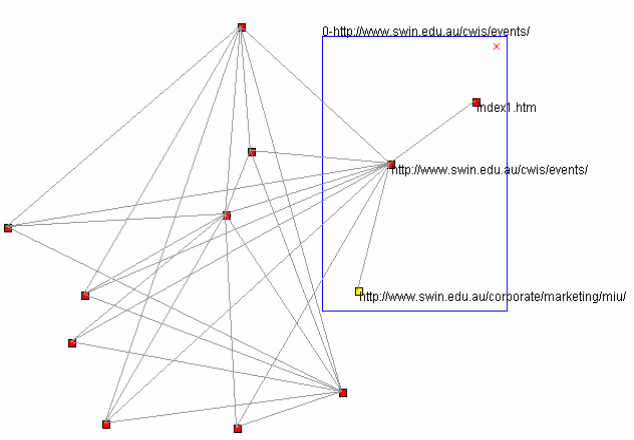


Fig. 13. Filter applied to the expanded node

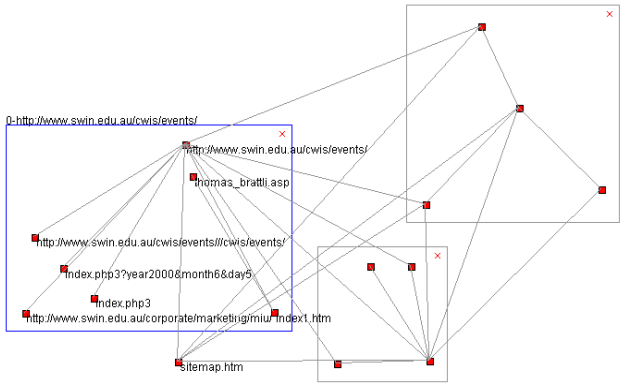


Fig.14. Expanded nodes

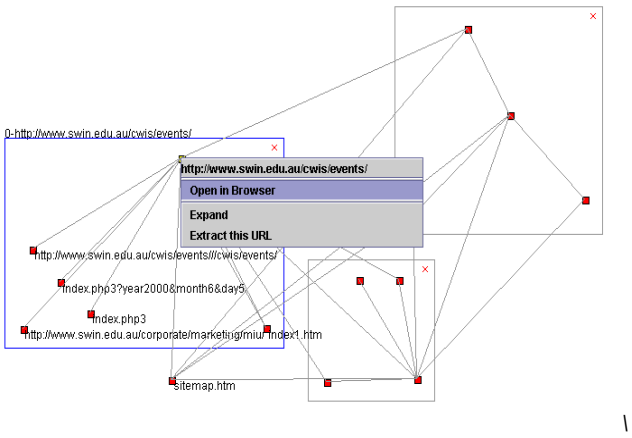


Fig.15. Node navigation menu

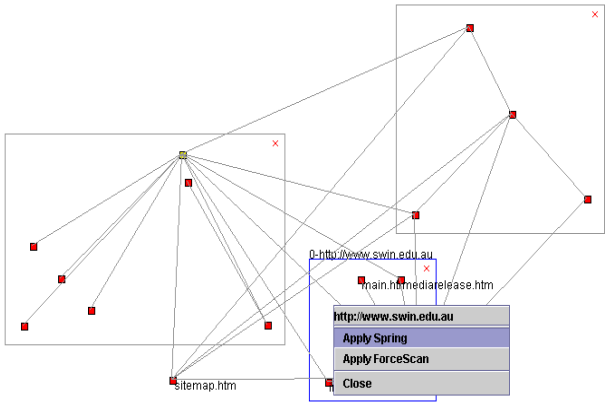


Fig.16. Navigation menu for expanded node

C. *Discussion*

Figure 11 shows the visualization result generated by FCG for the Website of Swinburne University with two levels of exploration depth. In Figure 11 graph drawing algorithms are applied to the Web graph, which assign the positions for all nodes to ensure that there are no overlapping nodes in the graph.

The running time taken by FCG to render the Web graph in Figure 11 is 1.3 seconds in Pentium 4 computer with 1.8GHz clock and 512 RAM. The drawing process using modified spring algorithm [5], took with the spring animation off. The modified spring algorithm only applied in fifty iterations, so if the animation to position the node is turned on, it took half second per iteration (the system timer in FCG to call the spring algorithm is per half second). The force-scan algorithm [10], used to remove the overlapping nodes, took 0.17 seconds to perform on ten nodes.

The Web crawler and clustering/filtering process took longer time than the layout process. The reason for this is that it depends on the Internet speed. The clustering process depends heavily on the number of URLs, taking 2.4 seconds to cluster 178 URLs in this case.

The ten nodes in the Web graph in Figure 11 can be shown in more detail as shown in Figure 12 by expanding the nodes. In Figure 12 the node http://www.swin.edu.au/cwis/events/ is expanded from the corresponding node in Figure 11, and graph drawing algorithms are applied on the expanded node window. Without animation, it took less then 1 second to apply the modified spring algorithm to the expanded node, and 0.22 seconds to apply the force-scan algorithm to the expanded node and root level nodes. The node expanded creates a new window, so it is easy to see the nodes inside the expanded node. When the window is closed, the Web graph in Figure 12 is returned back to Web graph in Figure 11.

Figure 13 shows the resulting graph when applied filtering to the Web graph in Figure 12. The weak links and unimportant link filters have been removed. The node labels in Figure 13 are visible only for the active window, in which the expanded node resides.

An example of the Web graph with more than one expanded node is illustrated in Figure 14 Note that the filtering rules have been applied to the graphs shown in Figures 13 and 14. In the presence of more than one expanded nodes, the Web graph tends to grow too large to be fitted in the screen. When this problem occurs, expanded nodes that are not focused and positioned outside will be closed.

The FCG system provides the navigation menu for a focus node, as shown in Figure 15. There are three menus: first, “Open in Browser”, which opens the URL page of the corresponding node using the system default browser; second “Expand” , which expands the node into the detailed nodes, similar to the node http://www.swin.edu.au/cwis/events/ in Figure 12, and the last menu called “Extract this URL” , which extracts all other URLs connected to the current URL in order to expand the Web graph. When the latter menu is clicked, the entire visualization process described before will be performed again. For the updated graph two windows will be generated, along with the first window displaying the original graph layout, and the second showing the newly extracted URLs.

Figure 16 shows the navigation menu that is available for the expanded node when the focused expanded node is right-clicked. There are three menus items. The first menu item displays the focused expanded node name (a URL). The second menu item, “Apply Spring”, applies the modified spring algorithm to lay out the expanded node that includes all nodes inside it. The other nodes will, however, not be affected. The second menu item, “Apply force-scan”, enforces the force scan algorithm to adjust the layout of the expanded node and its parent nodes. Although there are three expanded nodes in Figure 16, for example, the force-scan algorithm is restricted to the nodes inside the window titled http://www.swin.edu.au, and their root level nodes, if the “Apply force-scan” menu item is chosen for the expanded node. The last menu item “close” closes the expanded node.

VI. CONCLUSION

In this paper, we have presented a system for visualization of Web sites, along with the algorithm and approach for clustering and filtering graphs. As opposed to existing approaches that suffer from the limitation of the messy layouts of large graphs, our approach was designed to overcome this difficulty in a stepwise and refinement way by using clustering and filtering graphs. A prototype called FCG has been implemented to demonstrate the performance of our approaches with a case study. The future work will include the usability test of this system.

REFERENCES

[1] G. D. Battista, P. Eades, R. Tamassia, and T. Tollis, *Graph drawing: algorithms for the visualization of graphs,* Prentice Hall, 1999.

[2] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," In *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, Australia, April 1998.

[3] Y. Chen and E. Koutsofios, "WebCiao: a Website visualisation and tracking system," In *Proceedings of WebNet 97 Conference*, 1997.

[4] J. Cho, H. Garcia-Molina, and L. Page, "Efficient crawling through URL ordering," In *Proceedings of the Seventh International World Wide Web Conference*, pages 161–172, April 1998.

[5] P. Eades, "A heuristic for graph drawing," *Congressus Numerantium*, 42:149-1601984.

[6] M. Huang, P. Eades, and J. Wang, "On-line animated visualization of huge graphs using a modified spring algorithm," *Journal of Visual Languages and Computing*, vol. 9, no.6, pp. 623-645,1998

[7] X. Huang and W. Lai, "Automatic abstraction of graphs based on node similarity for graph visualization," In *Proceedings of The Fifteenth International Conference on Software Engineering and Knowledge Engineering*, pp167-173, San Francisco Bay, July 2003.

[8] X. Huang and W. Lai, "NodeRank: a new structure based approach to information filtering," In *Proceedings of the International Conference on Internet Computing*, pp167-173, Las Vegas, USA. 2003.

[9] W. Lai, M. Huang, Y. Zhang, and M. Toleman, "Web graph displays by defining visible and invisible subsets," In *Proceedings of AusWeb99 - the Fifth Australian Web Conference,* pp. 207-218, Ballina, NSW, April 1999.

[10] W. Lai, M. Huang, and J. Tanaka, "Fitting Web graphs in a display area with no overlaps for Web navigation," In *Proceedings of the International Conference on Internet Computing*, pp. 601-607, June, 2002.

[11] W. Lai and P. Eades, "Removing edge-node intersections in drawings of graphs," *Information Processing Letters*, vol.81, pp105-110, 2002.

[12] Y. S. Maarek and I. Z. B. Shaul, "WebCutter: a system for dynamic and tailorable site mapping," In *Proceedings of the Sixth International World Wide Web Conference*, pp. 713-722, 1997.

[13] R. C. Miller and K. Bharat, "SPHINX: A framework for creating personal, site-specific Web crawlers," In *Proceedings of the Seventh International World Wide Web Conference*, pp.119–130, April 1998.

[14] K. Misue, P. Eades, W. Lai, and K. Sugiyama, "Layout adjustment and the mental map," *Journal of Visual Languages and Computing*, No. 6, pp. 183- 210

[15] C. Pilgrim and Y. Leung, "Applying bifocal displays to enhance WWW navigation," In *Proceedings of the Second Australian World Wide Web Conference*, 1996.

Association-Based Segmentation for Chinese-Crossed Query Expansion

Chengqi Zhang, Zhenxing Qin, Xiaowei Yan

***Abstract*—The continually and high-rate growth of China's economy has attracted more and more international investors. These investors have an urgent need of identifying patterns in Chinese information, which are potentially useful in making competitive decisions. The first step of deeply understanding and analyzing Chinese information is how to effectively search those likely relevant to a user query. However, queries provided by users are often incomplete and inappropriate to the information systems, especially for retrieving Chinese-crossed information. In this paper, we present a segmentation based on actionable Chinese term-association analysis for better understanding user queries so as to automatically generate Chinese-crossed-query expansions. The semantics behind the actionable term-association rules is thus studied. Experiments conducted have shown that our approach is efficient and promising.**

***Index Terms*—Chinese and Japanese characters, Information retrieval, Text processing,**

I. Introduction

The pressure of enhancing corporate profitability has caused companies to spend more time on identifying diverse sales and investment opportunities for winning China's markets. The short list of examples below should be enough to place the current situation into perspective (these examples are cited from [12]):

- The year 2008 may seem like a long way away, but commercial enterprises know that now is the time to line up for sponsorship of Beijing's Olympic Games. It's the moment China has been dreaming about for years. It is a time of national pride and celebration and also a time of opportunity. Many international companies see the 2008 Olympics as a chance to earn huge profits. (see http://www.creadersnet.com/newsViewer.php?idx=99079)
- Credit Suisse-First Boston (CSFB) says it has listed the China market as one of its primary targets for business development and is focusing its global resources on the country. John J. Mack, CEO of the CSFB, told that China's continued economic reforms and its entry into the World Trade Organization have fostered great potential for business development. (see http://www.creadersnet.com/newsViewer.php?idx=129913)
- PeopleSoft, the fifth largest software maker in the world, will invest in China in the coming eight years. "We make a few investments every year and China will be our biggest move this year in terms of market and products," Craig Conway, chief executive officer and president of PeopleSoft. He said China's accession to the World Trade Organization (WTO) means more local enterprises will have chances to compete with their international counterparts and are anxious to upgrade their management with enterprise software, so PeopleSoft is coming at a "prefect time". (see http://www.creadersnet.com/newsViewer.php?idx=129912)

With the increasing interest to China's markets, the need to be able to digest the large volumes of Chinese information is now critical. In particular, it is very important to discover and develop Chinese ancient cultures and medicines, which brings benefit to mankind. The first step of deeply understanding and analyzing these Chinese information is how to effectively search those likely relevant to a user query. Accordingly, this paper focuses on the issues of Chinese information retrieval.

User queries to the Web or other information systems are commonly described by using one or more terms as keywords to retrieve information. Some queries might be appropriately given by experienced and knowledgeable users, while others might not be good enough to ensure that those returned results are what the users want. Some users consider that Boolean logic statements are too complicated to be used. Usually, users are not experts in the area in which the information is searched. Therefore, they might lack the domain-specific vocabulary and the author's preferences of terms used to build the information system. They consequently start searching with generic words to describe the information to be searched for. Sometimes, users are even unsure of what they exactly need in the retrieval. All of these reasons then often lead to uses of incomplete and inaccurate terms for searching. Thus, an information retrieval system should provide tools to automatically help users to develop their search descriptions that match both the need of the user and the writing style of the authors.

One of the solutions to provide the service is the automatic expansion of the queries with some additional terms [1, 2]. These expanded terms for a given query should be semantically or statistically associated with terms in the original query.

Manuscript received May 9, 2004. This work is partially supported by a large grant from Australian ARC (DP0559536), a China NSFC major research Program (60496321), and a China NSFC grant (60463003)

Chengqi Zhang is with the Faculty of Information Technology, University of Technology Sydney, PO Box 123, Broadway, Sydney NSW 2007, Australia. (e-mail: Chengqi@it.uts.edu.au)

Zhenxing Qin is with the Faculty of Information Technology, University of Technology Sydney, PO Box 123, Broadway, Sydney NSW 2007, Australia. (e-mail: zqin@it.uts.edu.au).

Xiaowei Yan is with Department of CS, Guangxi Normal University, Gulin, China. (e-mail: yanxw@mailbox.gxnu.edu.cn)

Moreover, techniques of association rule mining [9, 10] are frequently used for text mining [3, 5, 6, 8] and global query expansion [4, 7].

For Chinese information, users often search for information with Chinese-crossed queries. Because of the complicated morphology, syntax and semantics of Chinese language, it is very difficult to generate the efficient and intelligent service of Chinese-crossed text retrieval.

In this paper, we introduce our system for dealing with Chinese-crossed queries, in which a segmentation based on actionable Chinese Term-Association Rule (CTAR) analysis is proposed for better understanding user queries so as to automatically generate Chinese-crossed-query expansions. In Section 2, the system structure, the actionable CTAR analysis and the establishment of thesauri are depicted. In Section 3, performance of our method is evaluated based on our experiments.

II. Problem Statement

A. System Structure

Figure 1 illustrates the system structure that we design for Chinese-crossed query expansion.

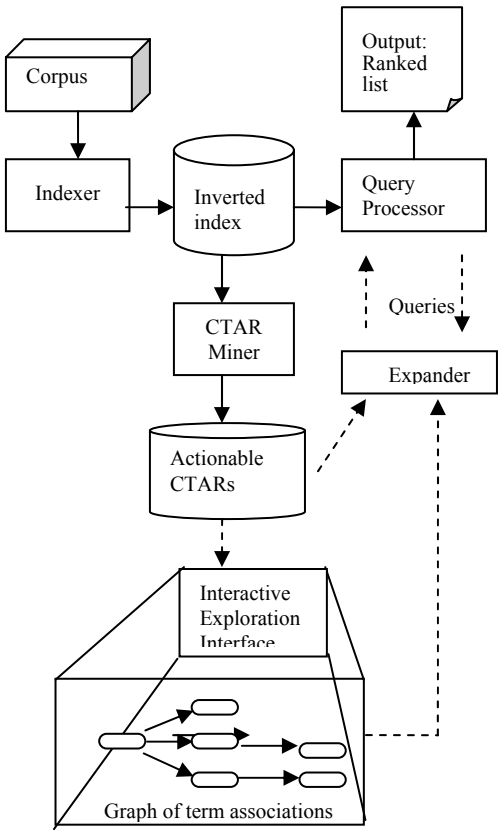


Figure 1: System structure for Chinese-crossed query expansion.

The system consists of four main parts: indexing module, actionable CTAR miner, query expander and query processor.

The indexing module constructs an inverted index for a corpus. Unlike English text processing, Chinese text processing does not need tokenizing and stemming, but a segmentation step is required to parse the Chinese text for a term list. We will discuss the segmentation later.

The actionable CTAR miner extracts those Chinese term-association rules actionable to Chinese-crossed query expansion. The graph of term-associations, thus constructed, can be interactively browsed by using the interactive exploration interface. User can control the confidence and support thresholds of actionable CTARs, as well as the frequency of the terms to be displayed, by browsing the graph of term-associations.

The query processor calculates a weighted-cosine similarity between a query and a document.

The query expansion module uses the term-association and Chinese Thesauri to expand queries.

B. The Segmentation Process

Chinese text is different from English text. Namely, there is no explicit word boundary in a sentence of Chinese. In English text, words are separated by spaces or punctuations. We can easily extract the terms from English text. In Chinese text, words are made up of one, two or more Chinese characters, and the same character can occur in many different words. The separators, such as comma and period, are only found between sentences. Furthermore, there is no explicit indication to tell where one word begins or ends in a Chinese sentence.

Decomposing a Chinese sentence into many single Chinese characters is not a good approach of segmentation. In Chinese, some words are composed of the same characters, but have different meanings in different sequences. Query expansion is a useful approach to improve the recall of retrieval. No matter whether you are expanding queries with a thesaurus or doing global/local analysis, you should work on substantive words. To apply the expansion methods to Chinese text retrieval, we need extract the words from sentences. How to exact words from Chinese text is always a challenge task in Chinese information retrieval. The process of breaking sentences into words is called segmentation.

There are two major segmentation techniques. One is the statistical approach and another is the dictionary-based approaches. The first method works well in finding bigrams. But its limitation is that it can only deal with words not longer than 2 characters. In our expansion system, the second segmentation method is chosen. It uses a lexicon tool (a Chinese wordlist) to find the word boundaries. It is more flexible and proper for query expansion.

The segmentation method exactly used in our system is called the forward maximum matching method. The segmentation is a Marlkov process, because the next word from segmentation is only decided by the current sentence and the dictionary, being unrelated with the words segmented before.

We first get a set of sentences from a Chinese text using the nature segmentation symbols, such as spaces, punctuations. For each sentence, we use forward maximum matching method with a dictionary finding the boundaries between words.

The dictionary we used contains 44,000 words with the length varying from 1 to 7 characters. We first put these words into 7 catalogues according to their length and sorted them meanwhile. The sorting is to facilitate the later searching. They can be read into memory during the segmentation process. These 7 catalogs are named wordlist1 to wordlist7 according to the length of the words inside the catalogues. The segment algorithm is shown below.

Segment (*sentence*)
Input: *sentence*, *wordlist1, … , wordlist7*
Output: a list of words

Step 0. word_list ← empty ;
Step 1. *l* ← length of the sentence
Step 2. **if** *l* = 0 **then** stop and output word_list;
Step 3. **if** *l* > 7, *l* ← 7;
prefix ← sentence (0, *l*)

Step 4. **if** prefix can be found in wordlist*l* **then**
{ add prefix to word_list;
sentence ← sentence - prefix;
goto step1}
else if *l*>2 { *l*← *l* –1; **goto** Step3}
else { prefix ← sentence (0,1);
add prefix to the output wordlist;
sentence ← sentence-prefix;
goto step1; }

C. Actionable Chinese Term-Association Rules

We set the minimum support at 0.0001 and minimum confidence at 0.1 in the rules mining. These thresholds are smaller than those for English corpus, because the Chinese words are distributed more evenly in different documents. Figure 2 shows the distribution of words frequency for a Chinese corpus introduced later. The horizontal axis represents document frequency in e~k. The vertical axis is the ratio of words, which fall into the document frequency. From the most-left point in Figure 2, we can see that about 10% of words occur in less than e1 documents. In AP90 from TREC4, the collection of news wires issued by the Associated Press in 1990 [7], there are 45% of words fall in this field. In this English corpus, we eliminate the stop words in indexing process, and there are still 455 words appearing in at least 5000 documents. In the Chinese corpus, on the other hand, we do not pick out stop words, and there are only 300 words appearing in more than 5000 documents. Because of this distribution over word frequency, there are less words co-appearing in the same documents. Therefore, we choose a smaller confidence threshold in Chinese CTARs mining, and later in rules selection in query expansion.

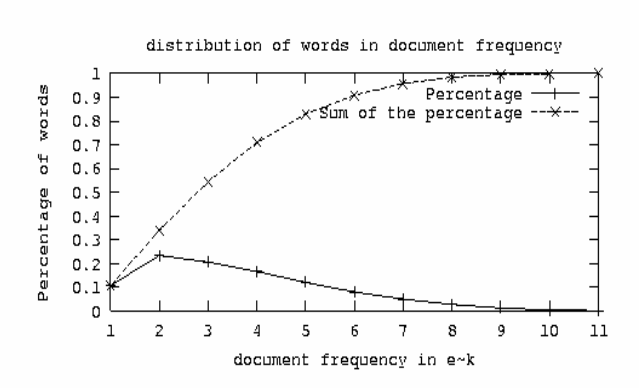


Figure 2: distribution of words frequency for a Chinese corpus.

D. Natural Semantics behind Actionable CTARs

A high confidence rule of the form t1 => t2 indicates that t2 often appears in a document if t1 appears. This suggests a certain type of relations between the terms, such as hypernym/hyponym or holonym/meronym, indicating a narrower/broader meaning between the terms. These relations characterize what we could call a contextual holonony, that is, if t1 => t2, then t1 is part of the vocabulary in the topical context, which is suggested by the concept denoted by t2. We categorized such rules into following four classes. The categorization is motivated from our inspection. Although this classification does not always seem reasonable, it can help us to understand the latent semantics behind those rules and their effects on the later query expansion.

(1) Hypernym/Hyponym ("a kind of" relation)

"Weltanschauung" is a kind of "ideology". "Balance beam" is a kind of "gymnastics". And "Marxism" is a kind of "theory". The corresponding rules are

世界观 (weltanschauung) => 思想 (ideology)

SUPP = .00089, CONF = .82258

平衡木 (balance beam) => 体操 (gymnastics)

SUPP = .00027, CONF = .8421

马克思主义 (Marxism) => 理论 (theory)

SUPP = .00125, CONF = .8372

(2) Meronym/Holonym ("a part of" relation)

"Bremen" is a part of Germany. "Ophthalmology" is a part of the hospital. "Wing" is a part of a aeroplane.

不莱梅 (Bremen) => 德国 (Germany)

SUPP = .00017, CONF = .90909

眼科 (ophthalmology) => 医院 (hospital)

SUPP = .00012, CONF = .875

机翼 (aerofoil) => 飞机 (aeroplane)

SUPP = .00034, CONF = 1

(3) Other narrower/broader meanings in topics

Sometimes, words in the rules are not strictly in hypernym/hyponym or meronym/holonym relationship. But they indicate a relationship of narrower/broader meaning in topic. "Sentence" is a narrower topic than that of "law". The other example is shown by cancer and hospital.

量刑 (sentence) => 法律 (law)

SUPP = .00015, CONF = .81818

羊城 (city of sheep) => 广州 (GuangZhou)

SUPP = .00017, CONF = .83333

侨胞 (emigrant) => 海外 (oversea)

SUPP = .00122, CONF = .85365

肿瘤 (cancer) => 医院 (hospital)

SUPP = .00013, CONF = .88888

(4) Special word usage in Chinese

In Chinese, there are some special word usages in verb-object group, subject-verb group and adjective-noun group. In verb-object group, for example, when mentioning a certain objects we mostly uses the corresponding verb, or (not and) vice versa. An example is that when an ambassador handover credentials, the verb for handover in Chinese must be "递交".

For subject-verb groups there is a similar situation. If "喝采" (applause) occurs, the subject is likely "观众" (audience).

Some adjectives are designated to modify specific nouns. For instance, "翻天覆地" (turn over the world) is specific for "变化" (change). "悠久" (age-old) is mostly used to modify "历史" (history). The related rules are as follow.

国书 (credential) => 递交 (handover)

SUPP = .00031, CONF = .81818

喝采 (applause) => 观众 (audience)

SUPP = .00022, CONF = .8125

决口 (breach) => 洪水 (flood)

SUPP = .00027, CONF = .8421

翻天覆地 (turn over the world) => 变化(change)

SUPP = .00076, CONF = .93617

悠久 (age-old) => 历史 (history)

SUPP = .00055, CONF = .94117

A rule t1 <=> t2, i.e. t1 => t2 and t2 => t1 with high and similar respective confidences, conf1 and conf2 respectively, as well as a sufficient support, indicates that t1 and t2 tend to appear together. We refer to t1 and t2 as “context synonyms”. Several kinds of “context synonyms” are given below.

(1) Synonym

t1 and t2 are real synonyms. For example, "防汛" (flood prevention) and "抗洪" (fight a flood), "漏税" (evade taxation) and "偷税" (evade taxes), "腐朽" (rotten) and "侵蚀" (erode), "检举" (report an offense to the authorities) and "揭发" (expose a crime) are all synonymic pairs. Here also list some of the rules.

防汛 (flood prevention) <=> 抗洪 (fight a flood)

SUPP = .0008, CONF1 = .32857, CONF2 = .34586

断流 (a river stops flowing) <=> 干涸 (dry up)

SUPP = .00012, CONF1 = .46666, CONF2 = .4375

漏税 (evade taxation) <=> 偷税 (evade taxes)

SUPP = .00012, CONF1 = .36842, CONF2 = .30434

腐朽 (rotten) <=> 侵蚀 (erode)

SUPP = .00061, CONF1 = .47945, CONF2 = .59322

检举 (report offenses to authorities) <=> 揭发 (expose crimes)
SUPP = .00026, CONF1 = .42857, CONF2 = .6

姑息 (tolerate evil) <=> 迁就 (yield to)

SUPP = .00026, CONF1 = .42857, CONF2 = .68181

受贿 (accept bribes) <=> 贪污 (corruption)

SUPP = .00075, CONF1 = .46236, CONF2 = .42574

(2) Antonym

It is an interesting relation. In Chinese, antonyms are often used together to emphasize something, such as extensive and intensive, modulation and demodulation.

粗放 (extensive) <=> 集约 (intensive) (in farming or management)

SUPP = .0008, CONF1 = .34328, CONF2 = .33823

调制 (modulation) <=> 解调 (demodulation)

SUPP = .00043, CONF1 = .64102, CONF2 = .92592

男生 (schoolboy) <=> 女生 (schoolgirl)

SUPP = .00012, CONF1 = .63636, CONF2 = .41176

(3) Peers relation.

Some closely-related peers appear together, such as "springboard" and "platform" in diving, "uneven bars" and "balance beam" in gymnastics, "Franc" and "Pound" in currency, "Germany" and "France" in countries, etc.

跳板(跳水) (springboard diving) <=> 跳台(跳水) (platform diving)

SUPP = .00047, CONF1 = .58695, CONF2 = .4909

高低杠 (uneven bars) <=> 平衡木 (balance beam)

SUPP = .0002, CONF1 = .46153, CONF2 = .63157

法郎 (franc) <=> 英镑 (pound)

SUPP = .00323, CONF1 = .42923, CONF2 = .73122

德国 (Germany) <=> 法国 (France)

SUPP = .01813, CONF1 = .33977, CONF2 = .39694

(4) Country and its capital

Though capital is a part of a country and the real relation between them should be viewed as meronyms/holonyms. They often appear together in context. When an author mentions a country, its capital often appears in the document and vice versa. An example is that between 贝鲁特 (Beirut) and 黎巴嫩 (Lebanon).

黎巴嫩 (Lebanon) <=> 贝鲁特 (Beirut)

SUPP = .00234, CONF1 = .5214, CONF2 = .64734

泰国 (Thailand) <=> 曼谷 (Bankok)

SUPP = .00146, CONF1 = .37004, CONF2 = .50299

(5) People and their workplace

Some people work in certain places. For example, 民警 (policeman) works in 派出所 (local police station), 渔民 (fisher) works in渔船 (fishing vessel), etc. Please look at the rules below for more details.

民警 (policeman) <=> 派出所 (local police station)

SUPP = .00083, CONF1 = .48, CONF2 = .41379

渔民 (fisher) <=> 渔船 (fishing vessel)

SUPP = .00031, CONF1 = .54545, CONF2 = .52941

守门员 (goalkeeper) <=> 球门 (goal)

SUPP = .00043, CONF1 = .38461, CONF2 = .36231

农民 (farmer) <=> 农村 (countryside)

SUPP = .01193, CONF1 = .42554, CONF2 = .48234

(6) Special word usage in Chinese

We already mentioned some special word usage in Chinese previously. Some words in the group depend on each other and are often used together. A subject-verb example is 候选人 (candidate) and 竞选 (election contest). A verb-object example is 体察 (observe) and 民情 (condition of the people). There are some phrases also can be considered in this category, such as "不仅" (not only) and "而且" (but also), "通俗" (popular) and "易懂" (easy to understand).

候选人 (candidate) <=> 竞选 (election contest)

SUPP = .00207, CONF1 = .40067, CONF2 = .40202

体察 (observe) <=> 民情 (condition of the people)

SUPP = .00015, CONF1 = .5625, CONF2 = .40909

通俗 (popular) <=> 易懂 (easy to understand)

SUPP = .00027, CONF1 = .55172, CONF2 = .8421

不仅 (not only) <=> 而且 (but also)

SUPP = .03548, CONF1 = .53098, CONF2 = .60464

(7) Local term-associations

There are also some relations only held in the corpus like those in English corpus. 海峡 (strait) and 台湾 (Taiwan) is an example. In Chinese news, 海峡 (strait) is often referred to Taiwan Strait and vice-versa.

海峡 (strait) <=> 台湾 (Taiwan)

SUPP = .00365, CONF1 = .50483, CONF2 = .34431

Examples of rules shown above have relatively high confidence, but not necessarily. Some rules with lower confidences (≈ 0.1) are still meaningful. For example,

结婚 (marry) <=> 婚礼 (wedding)

SUPP = .00015, CONF1 = .3, CONF2 = .10227

III. Performance of Chinese Query Expansion with Actionable CTARs

After exploring the semantic meaning of actionable CTARs in last section, we now tried to use these rules in query expansion.

A. Corpus and Queries

The corpus is a collection of news from the Xinhua News Agency in 1990. It has 57,240 documents. We segment the texts into Chinese words by using a dictionary. We do not eliminate stop words by using a stop-list. The documents are indexed after segmentation. There are 39,122 distinct words appearing in this corpus. The length of documents ranges from 5 to 3558 words, with 321 words on average.

We use 10 queries from TREC5 Chinese queries in our experiment, as listed below. E-title is the English translation from Chinese query. C-title is the original Chinese query, being used in retrieval.

1. <E-title> Communist China's position on reunification

<C-title> 中共对于中国统一的立场

2. <E-title> The newly discovered oil fields in China.

<C-title> 中国大陆新发现的油田

3. <E-title> Regulations and Enforcement of Intellectual

Property Rights in China

<C-title> 中国有关知识产权的立法与政策以及执法情况

4. <E-title> Numeric Indicators of Earthquake Severity in Japan

<C-title> 地震在日本造成的损害与伤亡数据

5. <E-title> Drug Problems in China

<C-title> 中国毒品问题

6. <E-title> World Conference on Women

<C-title> 世界妇女大会

7. <E-title> The Debate of UN Sanctions Against Iraq

<C-title> 联合国对伊拉克经济制裁的辩论

8. <E-title> The Mid-East Peace Talks

<C-title> 中东和平会议

9. <E-title> Measures to Prevent Forest Fires in China

<C-title> 中国森林火灾的防范措施

10.<E-title> Robotics Research in China

<C-title> 中国在机器人方面的研制

Because the corpus we use is not from TREC5, we can not obtain answers for the above queries from TREC5. It is impossible for us to read all the news and find out the exact precisions and recalls. Instead, we retrieve the top m documents for each query, and estimate the relevance of these documents. We then compare the average precision of these top m documents.

B. Expansion Methods and Retrieval Model Used

We tried three expansion methods based on three kinds of directions of rules, namely query words to expanded words, expanded words to query words, and dual directions.

We choose the parameters of support and confidence and varied them for each method, and select the best expanded terms (the expanded words most related to the queries). Then we retrieved the documents by four kinds of queries, the original query and the queries expanded with the three expansion methods.

The retrieval model is the Vector Space Model. We use the same similarity formula as that in [3]. The returned documents are sorted by their similarity values in descending order. We only consider the top m documents, ranked from 1 to m.

C. Relevance Estimation and Precision Value Computation

As discussed before, we retrieve documents by using a query and their three expanded queries respectively. The first m documents in the ranking list are selected for each query. Hence, we have at most 4*m documents for each query. Actually, there are many documents are retrieved, but only no more than 4*m documents are chosen for estimation for each query. In our experiment, when m = 10, there are 17.9 documents on average to be read for each query.

After retrieval, the retrieved documents are sent to a group of persons for estimation of the relevance to the query. The group of readers are all Chinese native speakers, but not experts on information retrieval and. We use the readers' votes as the value of precision. If there are n users who mark the first-ranked document relevant to a query, we say that precision of the document is n/10. The average precision of a query is the average precision of m documents for the query. The average precision of a rank is the average precision of the documents with the rank for all the 10 queries. The performance of average precision on each rank is shown below.

D. Comparison of Expansion Methods

We now compare the three expanded queries with the original query respectively. The query of No. 9, “中国森林火灾的防范措施” (Measures to Prevent Forest Fires in China), is used to illustrate our comparison.

(1) Expansion with Rules Q=>X

The expanded terms are relatively common words with high frequency. The average document frequency is about 4700. From the distribution of words in document frequency as shown in Figure 1, a word with this frequency (between e8 and e9) is among the top 5% frequent words.

Although we choose a small confidence of 0.2, it's still difficult to expand query terms which have high document frequency. This method favours those query terms with relatively lower document frequency, that is, those queries are easy to be expanded with general meaning words.

For the query of "Measures to Prevent Forest Fires in China", the expanded words are 生态(ecology), 保护 (protect), 环境 (environment), 林业 (forestry), 面积 (area), 检查 (inspect), 安全 (safety), 犯罪 (crime), 案件 (law case), etc. The corresponding rules are

森林 (forest) => 生态 (ecology), 保护 (protect), 环境 (environment), 林业 (forestry), 面积 (area)

火灾 (fire disaster) => 检查 (inspect), 安全 (safety), etc.

防范 (prevent) => 犯罪 (crime), 案件 (law case)

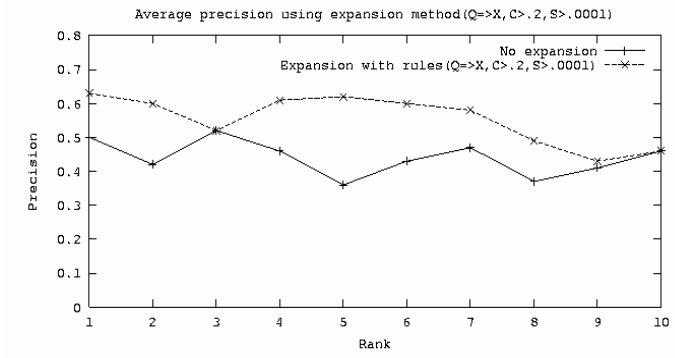


Figure 3: The average precision for each rank with rules Q=>X

The average precision over the ten ranks for this query is improved from 0.45 to 0.78 or 73.33%. Similarly, we expand all the 10 queries. The improvement in average precision is 25.91% (from 0.44 to 0.554). Performance of the query No. 2 is a little worse than that of original query. Performance of query No. 3 gains less than 5% of improvement. Query No. 5 has more than 10% improvement in performance. The expansions improve the precision in all ranks, except the third and tenth ranks where the performance keeps the same. Figure 3 shows the average precision for each rank.

(2) Expansion with Rules Q<=X

This method expands the query with words which imply occurrence of the query terms. On the contrary to the method of Q=>X, this method favours the relatively common query terms which have high document frequency, and expand the original query with specific meaning words.

We choose a higher confidence and support here to reduce the amount of expanded words and increase the quality of those words. For the query of No. 9, only one word is expanded, i.e. 防火 (fire prevention). The rule is

防火 (fire prevention) => 火灾 (fire disaster).

After expansion, the average precision is promoted from 0.45 to 0.67 or 48.89%. Because of the strict choose of parameters, only 8 queries out of 10 are expanded. When computing the overall average precision, we use the precision of original query for the two unexpanded queries. Improvement of the overall average precision is 27.95% (from 0.44 to 0.563). We only obtain significant improvement in the performances with two queries, but no decreased performance is found.

When look into the 10 ranks, there is only a little decrease at rank 6th. The other ranks all get benefit. See Figure 4.

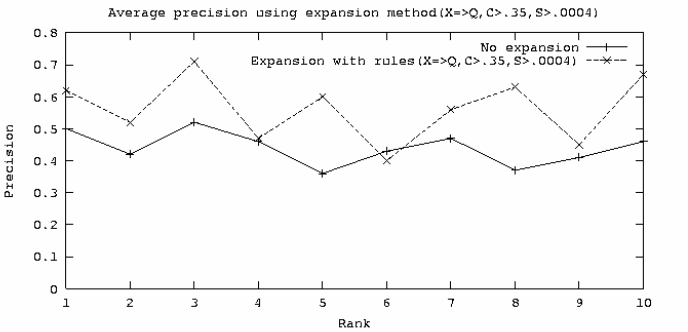


Figure 4: The average precision for each rank with rules Q<=X.

(3) Expansion with Rules Q<=>X

We choose a lower confidence of 0.1 and support of 0.0001 here, since there are not many such rules with high confidences in both directions. The rules for the query of No. 9 are

森林 (forest) <=> 森林资源 (forest resource), 生态 (ecology), 林区 (forest area),林业 (forestry);

火灾 (fire disaster) <=> 消防 (fire control/prevention/fighter), 大火 (conflagration), 火势 (fire impetus), 防火 (fire prevention).

After expansion, the average precision was promoted from 0.45 to 0.78 or 73.33%. This method expands all the 10 queries. Improvement of the overall average precision is 14.77% (from 0.44 to 0.505). We get a significant improvement in performance with the query of No. 3, but there are three queries with which a decreased performance is found as well. Moreover, this method shows inefficient with rank 1 and rank 9, and beneficial with other ranks. This is the worst among the three expansion methods. The performance of 10 ranks is shown in Figure 5.

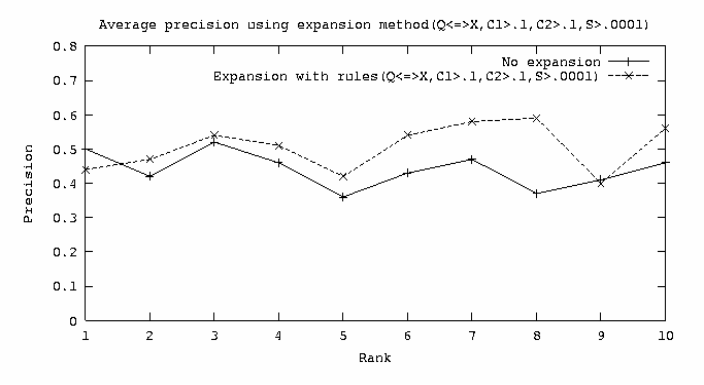


Figure 5: The average precision for each rank with rules Q<=>X.

IV. CONCLUSIONS AND FUTURE WORK

Recognizing the importance of Chinese information, we have developed a system for automatically generating Chinese-crossed-query expansions with CTAR mining techniques. The most feature of our system is to segment Chinese information into clusters such that each cluster contains those words that have a same character. We have demonstrated that our method can improve the effectiveness of retrieval over the corpus from the Xinhua News Agency. It is quite natural and reasonable to apply correlations among terms to the query expansion. Therefore, we believe that the idea in this paper is promising.

In the future work, we still need to study the thresholds determination for the actionable CTAR mining. Since there are many expansion methods, we should consider how to integrate these methods.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive comments on the first version of this paper. These comments have contributed to a vast improvement of this paper.
Also, we would like to thank Jie Wei and Dr. Shichao Zhang for their cooperative and discussion work for this paper.

REFERENCES

[1] Efthimis N. Efthimiadis. Query expansion. In: Martha E. Williams edited, *Annual Review of Information Scienc and Technology* (ARIST), Volume 31, pages 121-187, 1996.

[2] Mathias Géry and M. Hatem Haddad. Knowledge discovery for automatic query expansion on the World-Wide Web. In: Proceedings of Advances in Conceptual Modeling: ER '99 Workshops, Lecture Notes in Computer Science 1727, Springer, pages 334-347, Paris, France, November 15-18, 1999.

[3] Hatem Haddad, J.P. Chevallet, M.F. Bruandet. Relations between Terms Discovered by Association Rules. In: Proceedings of the 4th European Conference on Principles and Practices of Knowledge Discovery in Databases PKDD'2000, Workshop on Machine Learning and Textual Information Access, Lyon France, September 12, 2000.

[4] Jie Wei, Stéphane Bressan, Beng Chin Ooi. Mining Term Association Rules for Automatic Global Query Expansion: Methodology and Preliminary Results. Proceedings of First International Conference on Web Information Systems Engineering (WISE'00)-Volume 1

[5] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler, O. Zamir. Text Mining at the Term Level. In Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98), pages 65-73, Nantes, France. September 1998.

[6] Maria-Luiza Antonie, Osmar R. Zaïane. Text Document Categorization by Term Association. In Proceedings of 2002 IEEE International Conference on Data Mining (ICDM'02), pages 19-26, Maebashi City, Japan, December 09 - 12, 2002.

[7] Jie Wei, Zhenxing Qin, Stéphane Bressan, Beng Chin Ooi. Mining Term Association Rules for Automatic Global Query Expansion: A Case Study with Topic 202 from TREC4. In Proceedings of Americas Conference on Information Systems 2000.

[8] Xiaowei Yan, Chengqi Zhang, Shichao Zhang: Identifying Frequent Terms in Text Databases by Association Semantics. In: Proceedings of 2003 International Symposium on Information Technology (ITCC 2003), 28-30 April 2003, Las Vegas, NV, USA. IEEE Computer Society 2003: 672-675

[9] Chengqi Zhang and Shichao Zhang, Association Rules Mining: Models and Algorithms. Springer-Verlag Publishers in Lecture Notes on Computer Science, Volume 2307, p. 243, 2002.

[10] Shichao Zhang, Chengqi Zhang and Xiaowei Yan, PostMining: Maintenance of Association Rules by Weighting. Information Systems, Volume 28, Issue 7, October 2003: 691-707.

[11] "WordNet - a Lexical Database for English", www.cogsci.princeton.edu/~wn/

[12] http://www.creadersnet.com/newsPool/.

A Partial-Repeatability Approach to Data Mining

Kai-Yuan Cai, Yunfei Yin, and Shichao Zhang *Department of Automatic Control, Beijing University of Aeronautics and Astronautics, Beijing , China*

***Abstract*—Unlike the data approached in traditional data mining activities, software data are featured with partial-repeatability or parepeatics, which is an invariant property that can neither be proved in mathematics nor validated to a high accuracy in physics, but still (partially) governs the behavior of the data. Parepeatics emerges as a result of the inaccurate universe. The universe comprises all possible C language programs is an example that cannot be accurately characterized since human writes defect-prone programs. In this paper we design a parepeatic mining framework for software data diming, where the mined knowledge is represented in terms of parepeatic models. A parepeatic model consists of central knowledge, a knowledge fluctuation zone and a correctness factor. Our approach can generate the required parepeatic model as a new form of knowledge representation from a given dataset and apply it to software data mining. Experimental results with real C language programs show that the proposed approach is effective.**

***Index Terms*—Knowledge representation, parepeatic model , parepeatics, Partial-repeatability, software data mining, uncertainty**

I. INTRODUCTION

Data mining is an active or vigorous area nowadays and has found extensive applications in market analysis, investment assessment, security guarantee, manufacturing process analysis, Web searching, and scientific data analysis, among others [1-4]. It attempts to reveal or discover valuable information or knowledge from vast amount of data. The valuable knowledge can be represented in terms of patterns, clusters, association rules, significant structures and so on. These forms of knowledge can be discovered using various approaches such as clustering techniques, decision trees, neural networks and case-based reasoning methods. In contrast with traditional statistical data analysis that is assumption-driven and applicable to analyzing modest amount of data, data mining is discovery-driven and applicable to analyzing vast amount of data.

This work was supported by the National Natural Science Foundation of China (60233020, 60474006, 60473067).

Kai-Yuan Cai is with the Department of Automatic Control, Beijing University of Aeronautics and Astronautics, Beijing 100083, China (corresponding author, phone/fax: +86-10-8231-7328; e-mail: kycai@buaa.edu.cn).

Shichao Zhang is an assistant president at the Guangxi Teachers University. He is also a senior research fellow at the Faculty of Information Technology, University of Technology, Sydney. Contact him at the Univ. of Technology, Sydney, Broadway NSW 2007, Australia (e-mail: Zhangsc@it.uts.edu.au).

Roughly speaking, there are two implicit assumptions underlying existing approaches for data mining. First, there are invariant valuable patterns or knowledge among the dataset under mining. Second, the intended knowledge can be represented in a conventional form such as equivalence classes (clusters), statistical models, decision trees, induction rules, and neural networks. The first assumption is concerned with technical aspects as well as non-technical aspects. From the technical viewpoint we need to consider if the intended knowledge can be mined from the given dataset or if the given dataset is minable. From the non-technical viewpoint we need to consider if the data mining process is cost-effective. The second assumption is mainly technical. For example, suppose the given dataset is $\{x_1, x_2, \ldots, x_n\}$ and we want to cluster them into a number of classes. Then existing clustering techniques [5] will generate a few disjoint classes of data, $\{C_1, C_2, \ldots, C_m\}$, from the dataset as the intended knowledge. The second assumption implies that $\{C_1, C_2, \ldots, C_m\}$ is an appropriate representation of the intended knowledge.

However, our previous work in software data analysis has revealed that second assumption mentioned above can hardly hold in software engineering as a result of partial-repeatability featured with software data unless new models of knowledge representation are introduced [6, 7]. For example, given a software program, we may calculate the number of lines of code, the number of distinct usages of operators, the number of distinct usages of operands, the number of total usages of operators, and the number of total usages of operands. Then are there any invariant laws that govern these five measures regardless of the particular features of various software programs? The laws, if any, can hardly be represented as a single absolute assertion such as a deterministic function $y = f(x)$ or a statistical model. A more appropriate representation form is the one that includes central function (knowledge), fluctuation zone, and the corresponding correctness factor [7]. Such a new model or representation form is intended to characterize the feature of partial-repeatability in vast amount of software data. By partial-repeatability it is meant that complex phenomena may demonstrate an invariant property that can neither be proved in mathematics nor validated to a high accuracy in physics, but still (partially) governs the behavior of the phenomena. (Conventional science and technology follows the top criterion of full repeatability that scientific arguments must either be proved repeatably in

mathematics or validated to a high accuracy repeatably (even in a statistical sense) in physics (experimentally)) This means, traditional data mining techniques are not appropriate to software data analysis/mining.

In this paper we design a parepeatic mining framework for software data mining for dealing with the partial-repeatability problem. The main contributions in this paper include:

The notion of partial-repeatability is further clarified as a new kind of uncertainty in comparison with randomness and fuzziness.

The parepeatic model[1] is treated as a new form of knowledge representation. This model comprises the central knowledge such as valuable patterns and association rules, a fluctuation zone, and a correctness factor. If the fluctuation zone contains only the central knowledge and the correctness factor is equal to one, then the parepeatic model reduces to an existing or conventional model of intended knowledge.

A new criterion is introduced to measure the quality of conventional clustering. This criterion takes account of not only the homogeneity within each cluster and separability between the distinct clusters, but also the number of distinct clusters.

A new approach is proposed for data mining, which generates a partial-repeatability or parepeatic model from the given dataset for the intended knowledge.

The rest of this paper is organized as follows. Section 2 clarifies the notion of partial-repeatability as a new kind of uncertainty. Section 3 presents the proposed approach for mining data with partial-repeatability. Section 4 applies the new approach to mining software data. Concluding remarks are contained in Section 5. The Appendix details our mining algorithm for generating the required parepeatic model of clustering from the given dataset.

II. Partial-Repeatability as a New Form of Uncertainty

As mentioned in Section 1, before data mining, one must decide what he or she wants to extract from the dataset under mining. The intended knowledge must be represented in an appropriate form. If no uncertainty is associated with the given dataset and what we want to extract from the dataset is a causality relation or deterministic function, then we can use classic or crisp sets to represent the relation or function. For example, suppose the given dataset is $\{\langle x_{i1}, y_{i1}, x_{i2}, y_{i2}, \ldots, x_{im}, y_{im}, z_i \rangle; i = 1, 2, \ldots, n\}$, where x_{ij} is the number of copies of books of type j sell at a book store of concern on the ith day, y_{ij} is the price of books of type j sell at the book store on the ith day, and z_i is the total income of books sell at the book store on the ith day. If we are interested in the causes of z_i being increasing or decreasing, or we are interested in whether z_i is a increasing function of x_{ij} or y_{ij}, then the answer is positive and crisp. There holds $z_i = \sum_{j=1}^{m} x_{ij} y_{ij}$. This relation is deterministic and gives all information for the intended knowledge. No uncertainty is associated with the answer or intended knowledge. It is sufficient for a crisp singleton set of deterministic function or relation to represent the intended knowledge.

However one may argue whether the crisp answers of this kind are really interesting or valuable. They look trivial. More often than not, we are interested in answers with uncertainty. For example, what is the underlying relation among $\langle x_{i1}, y_{i1}, x_{i2}, y_{i2}, \ldots, x_{im}, y_{im}, z_i \rangle$ and i ? Suppose the intended knowledge is represented in form of $f(i; x_{i1}, y_{i1}, x_{i2}, y_{i2}, \ldots, x_{im}, y_{im}, z_i) = 0$. Then how to determine or represent the intended relation f ? Obviously, uncertainty must be associated with f and the answer can not be deterministic. A natural framework to represent the underlying uncertainty is probabilistic or statistical. We may assume that f is a random function. If n or the number of days of concern becomes huge and the relation among $\langle x_{i1}, y_{i1}, x_{i2}, y_{i2}, \ldots, x_{im}, y_{im}, z_i \rangle$ and i looks over complicated, then we may exploit human experience or heuristic and assume that f is fuzzy function. No matter whatever models (deterministic, random, or fuzzy) are used to represent the intended knowledge, existing approaches for data mining assume that the intended knowledge is represented in terms of a single relation f. More specifically, the single relation can be a statistical model, a fuzzy model, a decision tree, an induction rule, a neural network and so on.

Now the problem is whether a single relation f is sufficient to characterize the intended knowledge. By adopting the single relation f, we implicitly follow the philosophy of full repeatability that have been treated as the top scientific criterion in thousands of years of history of conventional science and technology. By full repeatability it is meant that scientific arguments can either be proved repeatably in mathematics or validated to a high accuracy repeatably in physics (experimentally). The correctness of the scientific arguments should be independent of the investigators who present or prove them. It is the nature of full repeatability or high quantitative accuracy that enables existing physics or science to gain enduring respect and reputation. Existing approaches for data mining treat the intended knowledge as a conventional scientific argument that can be fully repeatable even in a statistical sense.

Unfortunately full repeatability may fall as observed in our

[1] Parepeatics is an abbreviated term for partial-repeatability, and accordingly, 'parepeatic' is the adjective form of 'parepeatics'.

previous work [6, 7]. For example, a course of fish may not be fully repeatable if cuisine is involved: no chef can make a course of fish twice at exactly identical sweetness or saltiness. The bones of a human face may be slowly evolving; the skins may be slightly faster evolving. The face pattern continues to evolve over age. One cannot assure that the face at two different time instants would be fully identical, although they might be essentially similar. Things of this kind are only partially repeatable. This can be further justified in software engineering. No human can write exactly the same program twice for a single software requirement specification. No software test process can be exactly repeated twice. The behavior of software systems and the software development process can not be fully repeatable. However different software systems produced from different development processes for a single requirement specification can work similarly and do not fail in most cases, although one is not quite sure how to measure or quantify "similarly" or "most" even in a statistical sense. Software systems and software development processes behave partially repeatably. For the example given at the beginning of this section, we can treat the underlying relation as a function of n, denoted as $f^{(n)}$. Then how can $f^{(1)}, f^{(2)}, \ldots, f^{(n)}, \ldots$ behave fully repeatably as a single relation f? In very complicated situations, can $f^{(1)}, f^{(2)}, \ldots, f^{(n)}, \ldots$ behave partially repatably and fluctuates among a number of typical relations? We argue in our previous work [6, 7] that there is something lying between full repeatability (conventional scientific arguments) and miracles (the unrepeatable). This is partial-repeatability by which it is meant that complex phenomena may demonstrate an invariant property that can neither be proved in mathematics nor validated to a high accuracy in physics, but still (partially) governs the behavior of the phenomena. Partial-repeatability is a new kind of uncertainty that is distinctly different from randomness and fuzziness. If we treat the deterministic physical laws as type I laws (where causality dominates), the statistical physical laws as type II laws (where randomness dominates), then we can treat the laws governing the phenomena of partial repeatability as type III laws. In quantitative terms, a single relation or formula describing a type III law is not valid to a high accuracy (actually, a number of relations should be given), and substitution operations of variables may lead to significant errors.

Simply, we can identify partial–repeatability as a new kind of uncertainty as follows. Suppose $U = \{u\}$ is the universe of discourse and can be accurately characterized. For example, U comprises all positive integers. Further, let A be an object of interest (e.g., integers greater than 6, integers around 9). A is a crisp set if we can absolutely assert $u \in A$ or $u \notin A$ for all $u \in U$. If in general $u \in A$ holds to some extent and $u \notin A$ to another extent simultaneously, then A is a fuzzy set. If for any $u \in A$ there must be $A = u$ or $A \neq u$, but we are sure which u leads to $A = u$ or $A \neq u$, then we say that A is a random variable. Randomness and fuzziness assume that the universe of discourse is characterized accurately. However in some circumstances it is nearly impossible to characterize the universe of discourse accurately. For example, suppose the universe of discourse comprises all C language programs. Since human writes defect-prone programs and defects are in various forms, it is impossible to describe all possible C language programs. Another example can be encountered when we need to extract invariant patterns from all possible human faces. How can all possible human faces be represented accurately if they constitute a single U? In this way uncertainty is associated with the universe of discourse. Partial-repeatability emerges as a result of uncertain universes, no matter whether the relation between the object of interest and the universe of discourse is crisp, fuzzy or random. The object of concern is accordingly referred to as a parepeatic object. An example parepeatic object is the statement that the number of distinct usages of operators is less than that of distinct usages of operands in a C language program. The statement can hardly be assessed in a fuzzy or statistical context since the universe of C language programs is not accurate.

For the book-selling example mentioned at the beginning of this section, the universe of discourse comprises all possible selling scenarios that may take place at the book-store. The numbers of copies of books of various types sell at the book store, the corresponding prices and the number of days of concern serve as the features or feature variables defined for the universe, and the given dataset $\{\langle x_{i1}, y_{i1}, x_{i2}, y_{i2}, \ldots, x_{im}, y_{im}, z_i \rangle; i = 1, 2, \ldots, n\}$ is an observation of the universe in terms of the feature variables. The interested object is just the intended knowledge that we want to extract from the observation for the universe in terms of the feature variables. If we can accurately describe all possible selling scenarios, then we can argue that the uncertainty associated with the intended knowledge acts as randomness, fuzziness or a mixture of them. Otherwise the intended knowledge should be a parepeatic object.

Therefore we can introduce a framework of parepeatic data mining as described in Figure 2.1. The universe of discourse is parepeatic. Notice that there are elements in the universe that can not be given accurately as a result of the uncertainty associated with the universe. For each given element of the universe, we define several features or feature variables, which will lead to various observations. Based on the features and the corresponding observations, we need to extract intended knowledge for the parepeatic universe. So, in such a data-mining framework, two fundamental questions must be addressed. First, how should the intended knowledge be represented? Crisp, random and fuzzy models are not enough. We need a new form of knowledge representation. This is the so-called parepeatic model that will be defined in Section 3. Second, how can the intended knowledge be extracted from the given observations? An approach will be proposed for this in Section 3 and detailed in the Appendix.

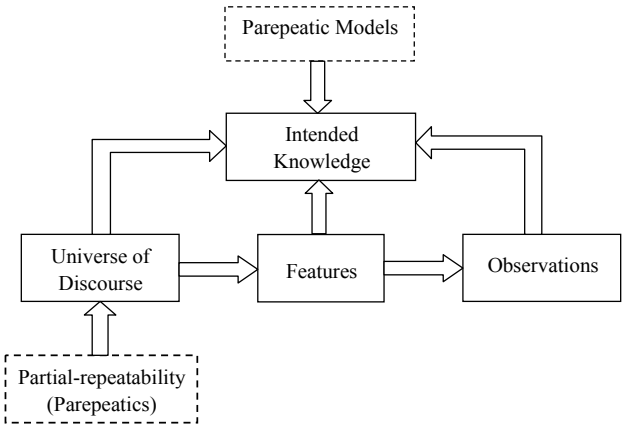


Figure 2.1 Parepeatic Data Mining

III. The Proposed Parepeatic Approach for Data Mining

A. Mathematical Definition of a Parepeatic Model

Suppose $U^{(p)}$ is the parepeatic universe of interest, and X the vector of feature variables of interest. For each element of $U^{(p)}$, X takes a vector value in $U^{(f)}$, which is referred to as the feature universe. Let f be a mapping from $2^{U^{(f)}}$ to V, where $2^{U^{(f)}}$ denotes the power set of $U^{(f)}$, and V is referred to as the knowledge universe of interest. f is the parepeatic object of interest. Suppose S is a subset of $U^{(f)}$ and defines the given set of values that X actually takes. Then we can write $f(S)$ as $f(X)$ if no confusion can otherwise arise. Note that $f(S)$ actually defines the central knowledge that we are going to extract from S for $U^{(p)}$ with the understanding that X is a simplifying representation of a generic element of $U^{(p)}$. Let $E(S)$ or $E(X)$ be a subset of V, or $E(X) \in 2^V$. We call the triplet $(f(X), E(X), c)$ a parepeatic model, where $f(X)$ is referred to as central knowledge[2], $E(X)$ the knowledge fluctuation zone, and c the corrector factor. There should hold $f(X) \in E(X) \subset V$ and $c \in [0,1]$.

In software engineering, we can use the parepeatic universe $U^{(p)}$ to represent the collection of all C language programs. Obviously, uncertainty is associated with the collection since it is nearly impossible to accurately define this collection. We have no idea how many C language programs there may be and whatever a C language program may be. This is particularly true if we consider the possibility that defects may be remaining in a C language program such that the synthetic and/or semantic requirements of C language are violated. Let $X = [x_1, x_2, \ldots, x_5]^\tau$ be the column vector of feature variables of interest, where τ denotes the transpose of a matrix and

x_1 : the number of lines of source code

x_2 : the number of distinct usages of operators

x_3 : the number of distinct usages of operands

x_4 : the number of total usages of operators

x_5 : the number of total usages of operands

X is a simplifying representation of a C language program, $X \in U^{(f)} = [0,\infty) \times [0,\infty) \times \cdots \times [0,\infty)$. Given a set of C language programs, suppose we are interested in clustering them. Then we can define $V = \{1, 2, 3, \ldots\}$. The central knowledge is $f(X)$ that defines the best number of clusters that the C language programs should be divided, the fluctuation zone $E(X)$ defines the set of all appropriate clustering, and c defines the degree of correctness of the clustering relation $f(X) \in E(X)$. Note that each cluster represents an equivalence class of C language programs in terms of X. Given a set of C language programs, there is a possibility that the given fluctuation zone $E(X)$ is inappropriate. For example, suppose the generated parepeatic model is $(3, \{2,3,5,7\}, 0.9)$. That is, the given set of C language programs suggests that all the C language programs should best be divided into 3 disjoint clusters, and it is also acceptable to divide all the C language programs into 2, 5, or 7 disjoint clusters. However these different clustering is still subject to uncertainty. One may argue that the best clustering that can be generated from the given set of C language programs should lead to $f(X) = 4$. Note $4 \notin \{2,3,5,7\}$. We only have $c = 0.9$ degree of confidence that $f(X) \in E(X)$.

B. Mining a Parepeatic Model from a Given Dataset

Suppose $\{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\}$ is a set of observations for the vector of feature variables X. We want to extract intended knowledge from the given dataset. Let the intended knowledge be represented in terms of a parepeatic model $(f(X), E(X), c)$. In general, the parepeatic model can be determined as follows.

Step 1. Obtain the dataset $\{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\}$ from n elements of the given parepeatic universe $U^{(p)}$.

Step 2. Pre-process the dataset $\{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\}$ to remove the undersirable data or outliers. Consequently, a new

[2] In our previous work [7] we call $f(X)$ the central function. Obviously function can be a form of knowledge of interest. However knowledge of interest can be defined in other forms such as clustering, patterns, association rules, and so on.

dataset, $\{X^{(1')}, X^{(2')}, \ldots, X^{(n')}\}$ with $n' \le n$, is obtained. The new dataset can be further transformed into $\{Y^{(1')}, Y^{(2')}, \ldots, Y^{(n')}\}$ if necessary.

Step 3. Determine the knowledge fluctuation zone $E(X)$ from the dataset $\{Y^{(1')}, Y^{(2')}, \ldots, Y^{(n')}\}$ according to some evaluation criterion.

Step 4. Determine the central knowledge $f(X)$ as a representative of $E(X)$ according to some evaluation criterion.

Step 5. Determine the correctness factor c finally.

In Step 3 a sampling technique is applied to sample a number of subsets of the data from the given dataset for the purpose of determining the knowledge fluctuation zone. These sampled data can be treated as the training dataset of the fluctuation zone. A sampling technique is applied also in Step 5 for the purpose of determining the correctness factor. The resulting data can be treated as the validation dataset of the fluctuation zone. Note that the above 5 steps for parepeatic data mining is rather abstract. The Appendix details how a parepeatic model as a new form of knowledge representation is generated from a given dataset, where a new criterion is introduced to measure the quality or performance of conventional data clustering in Step (11) of the algorithm detailed in the Appendix. The new criterion takes account of not only the homogeneity within each cluster and separability between the distinct clusters, but also the number of distinct clusters.

C. Discussion

One may observe that there is some similarity between a parepeatic model presented in Section 3.1 and a confidence interval representation in conventional statistical setting. However we should note that there are several essential differences between them. First, in conventional statistical interval estimation the parameter under estimation is given a priori. In a parepeatic model the perepeatic object of interest or the central knowledge $f(X)$ must be generated from the given dataset. It is not given a priori. A given dataset may generate different central knowledge, depending on the used generation criterion. Second, $f(X)$ is widely interpreted. It can represent clustering, patterns, scenarios, association rules and so on, depending on application context of interest. $f(X)$ is not stuck to a particular parameter. Finally, as we will observe more clearly in the rest of this paper, no statistical assumptions are taken for determining a parepeatic model. However statistical assumptions are essential for conventional interval estimations.

Notice the parepeatic data mining algorithm presented in Section 3.1 and the Appendix is closely related to data clustering. This is because we assume the knowledge universe characterizes the number of disjoint clusters of the parepeatic universe. That is, the central knowledge represents the number of clusters. However the algorithm presented in this paper differs from existing clustering algorithms in data mining [8] at least in two dimensions. First, existing data clustering algorithms measure the quality of clustering in terms of the homogeneity within each cluster and separability between the distinct clusters, but the number of distinct clusters is not taken into account. This is not true for the algorithm presented in this paper. Second, existing clustering algorithms generate a partition of the universe of discourse or the central knowledge. They do not produce the knowledge fluctuation zone or correctness factor.

In general, the essential difference between existing approaches to data mining and the parepeatic data mining approach lies in the different philosophies they follow and the different knowledge representation models they adopt. The existing approaches assume that the universe of discourse is accurately given and no partial-repeatability is involved. The adopted knowledge representation model is actually given in terms of central knowledge which can be clusters, statistical models, fuzzy models, decision trees, neural networks, induction rules and so on. On the other hand, the parepeatic approach assumes that the universe of discourse involves uncertainty and can not be given accurately. The adopted knowledge representation model comprises not only central knowledge, but also a knowledge fluctuation zone and a correctness factor. A parepeatic model coincides with an existing model if the knowledge zone contains only the central knowledge and the correctness factor is equal to one. The existing knowledge representation models can be treated as a special class of parepeatic models.

Here we note that data mining is related to the so-called granular computing [9]. Data clustering results in a number of disjoint clusters, which can each be treated as an information granule. Following this philosophy, we can see that there exists potential for the parepeatic data mining approach to be interpreted from the granular computing perspective.

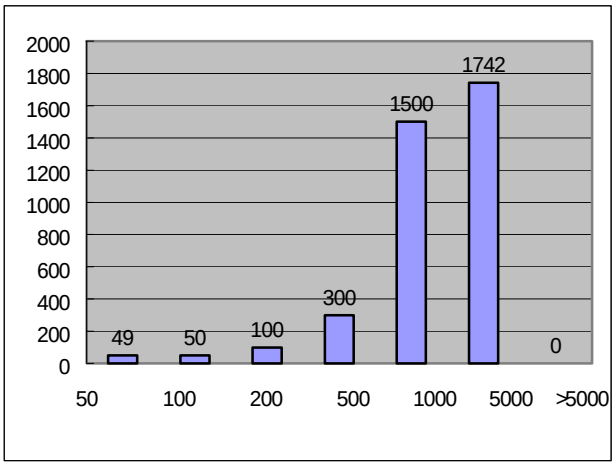
IV. Experimental Results in Software Data Mining

Software data mining, in some sense, can be traced back to Halstead's work on software science [10], although he neither adopted the terminology of data mining and nor recognized the importance of data mining. Halstead argued that there existed physics-like laws that obeyed each piece of software. He defined a number of software metrics such as those defined in Section 3 and proposed a set of the so-called software science formulae for these metrics. Empirical data were then collected from real software programs to validate the proposed formulae. We can treat these formulae as intended knowledge or intended laws and thus Halstead's work can be treated as a kind of software laws mining.

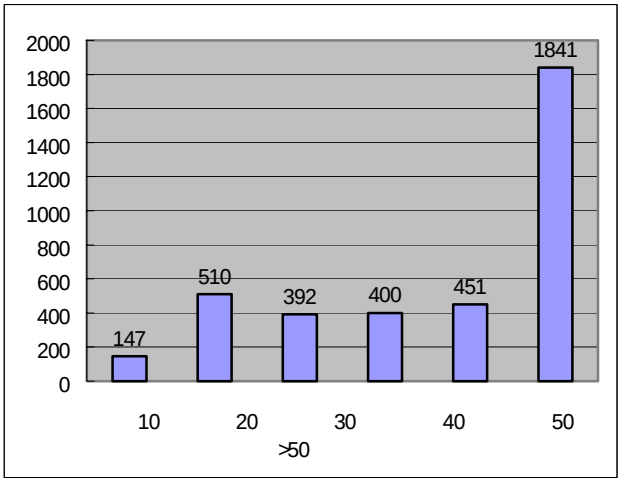
Unfortunately, as observed in our previous work [6, 7], Halstead's work still follows the philosophy of full repeatability and thus fails to stand. As argued in Section 2, the universe of software programs is parepeatic and thus software laws should be formulated in terms of parepeatic models. The

algorithm described in the Appendix can be employed to extract the required parepeatic models.

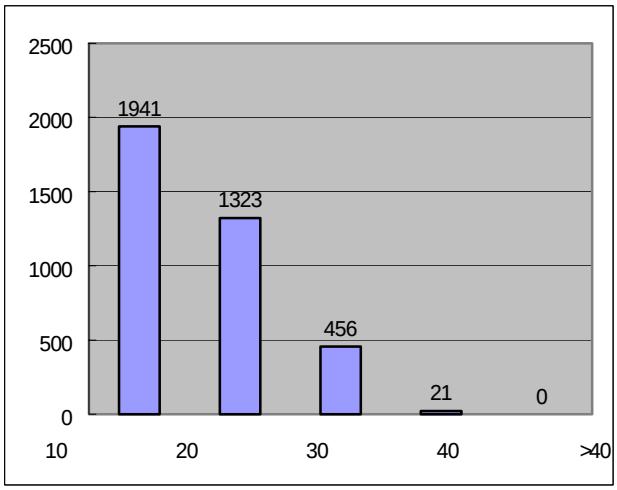
In order to test the algorithm described in the Appendix, we collected a set of 5437 C language programs. Some of the programs were downloaded from open source Web sites, and some of them came from undergraduate and graduate students' projects. For each program, we obtained a data point as that defined in Section 3. This has been performed in an automatic data collection tool. After the data pre-processing (i.e., Step (2) of the algorithm proposed in the Appendix), the dataset was reduced to a new dataset comprising 3471 data points. Figure 4.1 shows the histograms of these data points for x_1, x_2, x, x_4 and x_5. This new dataset ($\{T^{(1)}, T^{(2)}, \ldots, T^{(3471)}\}$ as specified in Step (2) in the algorithm described in Appendix) was then used to generate the required parepeatic model.



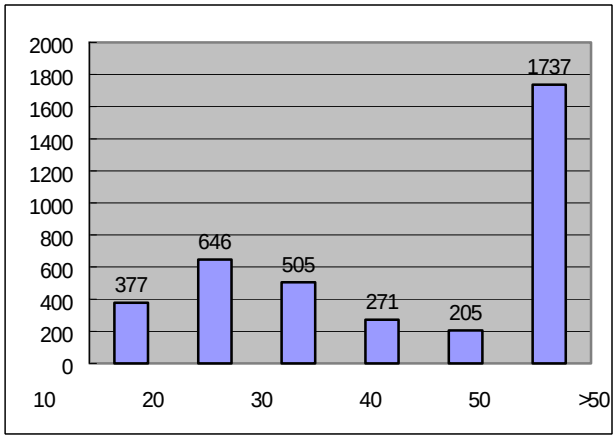
(a) Number of lines of source code



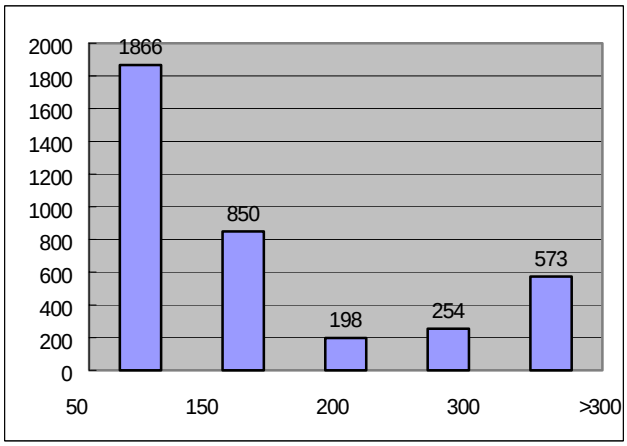
(b) Number of Distinct Usages of Operators



(c) Number of Distinct Usages of Operands



(d) Number of Total Usages of Operators



(e) Number of Total Usages of Operands

Figure 4.1 Histograms of Software Metrics for the Dataset

We carried out five experiments for the dataset $\{T^{(1)}, T^{(2)}, \ldots, T^{(3471)}\}$, that is, the algorithm described in the Appendix was applied for 5 times. In Steps (10) and (15) we applied a cut-set-based fuzzy clustering algorithm to do clustering, where the cut level was chosen as 0.975. Table 4.1 summarizes the experimental results. Different experiments

generated the training datasets of different sizes. However the validation datasets generated in these experiments had the same size ($\kappa_c \equiv 100$).

TABLE 4.1 EXPERIMENTAL RESULTS FOR PAREPEATIC SOFTWARE DATA MINING

($w_1 = 0.23, w_2 = 0.36, w_3 = 0.41, N = 25$)

Experiment No.	κ	Size of knowledge fluctuation zone	Central knowledge $f(X)$	κ_c	Correctness factor C	
					ε_0=0.001	ε_0=0.005
1	35	28	<11, 0.537088>	100	0.91	0.97
2	135	133	<11, 0.537075>	100	0.92	0.97
3	155	153	<10, 0.520748>	100	0.94	0.87
4	165	163	<10, 0.520748>	100	0.91	0.99
5	175	173	<10, 0.520748>	100	0.92	0.99

From the experimental results we can see that:

(1). As expected, greater value of ε_0 leads to greater value of C. This is understandable since Step (26) of the algorithm described in the Appendix implies that more data points are acceptable for the relation $f(X) \in E(X)$ if greater value of threshold is adopted.

(2). As the size of the training dataset grows, the central knowledge tends to be stable, although the size of the corresponding fluctuation zone tends to grow too. The corresponding correctness factor also behaves steadily. This implies that the generated parepeatic models are trustworthy and the algorithm described in the Appendix does work well.

V. CONCLUSION

Partial-repeatability or parepeatics emerges in complex phenomena if the complex phenomena demonstrate an invariant property that can neither be proved in mathematics nor validated to a high accuracy in physics, but still (partially) governs the behavior of the phenomena. This is particularly true in software engineering since software is developed by human. Software development processes and software system behavior are too complicated to be characterized accurately even in a statistical sense, but they still work and tend to serve human requirements, although they happen to fail to function. The notion of partial-repeatability was proposed in our previous work to contrast that of full repeatability that has been followed as the top criterion in traditional science and technology [6, 7]. By full repeatability it is meant that scientific arguments can either be proved repeatably in mathematics or be validated to a high accuracy repeatably (even in a statistical sense) in physics (experimentally).

In the preceding sections we have further clarified the notion of partial-repeatability as a new kind of uncertainty that is distinctly different from randomness and fuzziness. Suppose the universe of discourse is given and characterized accurately and an object is of interest. If the relation between the object and the universe is crisp, or each of the elements of the universe can be clearly identified to belong or not to belong to the object, then the relation is binary and the object is a crisp set. If the relation can be determined clearly, and each of the elements of the universe can belong to the object to some extent, and cannot belong to the object to another extent simultaneously, then the relation is fuzzy and the object is a fuzzy one. If each of the elements of the universe must either belong to or not belong to the object, but which elements belong to the object is not clearly determined, then the relation is random and the object is a random one. On the other hand, if the universe of discourse cannot be characterized accurately, then partial-repeatability emerges as a new kind of uncertainty. The corresponding object is a parepeatic one. An example universe is the one that comprises all possible C language programs.

Following the notion of partial-repeatability, we have proposed a new approach to data mining. In this approach the intended knowledge that is to be extracted from the given dataset is treated as a parepeatic object and represented in terms of parepeatic models. A parepeatic model consists of central knowledge, a knowledge fluctuation zone, and a correctness factor. Although a parepeatic model looks similar to conventional statistical confidence interval in some sense, there are essential differences between them. We have shown how to generate a parepeatic model (intended knowledge) from the given datset. The effectiveness of the proposed approach is justified by our experiments with software data mining.

The importance of the work presented in this paper is two-fold. First, partial-repeatability is clearly identified as a new kind of uncertainty that is distinctly different from randomness and fuzziness. This implies that we need to develop new mathematical framework to characterize this new kind of uncertainty and explore the underlying physical laws (or type III laws as mentioned in our previous work [7]). Second, a new framework of data mining, i.e., parepeatic data mining, is proposed. In this framework the intended knowledge is treated as a parepeatic object and parepeatic models are treated as a new form of knowledge representation. A lot of research work can be done by extending existing frameworks of data mining to the parepeatic counterpart. Pareaptic data mining reduces to conventional data mining if no uncertainty is associated with the underlying universe of discourse. This paper is only a small step towards to a new research direction and is speculative somewhat.

APPENDIX

Algorithm of Mining a Parepeatic Model:

In the context of software data mining, suppose the given parepeatic universe $U^{(p)}$ comprises all possible C language programs and the feature variables are $x_1, x_2, \ldots, x_5$ as given in Section 3.1. Given the observations

$\{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\}$, with $X^{(j)} = [x_1^{(j)}, x_2^{(j)}, \ldots, x_5^{(j)}]^\tau$, where τ denotes the transpose of a matrix (vector) we want to cluster them into several classes. After data pre-processing, we have $\{Y^{(1')}, Y^{(2')}, \ldots, Y^{(n')}\}$ and want to divide these data into a number of disjoint classes, say, $\{D_1, D_2, \ldots, D_m\}$, where $D_i = \{Y^{(i_1)}, \ldots, Y^{(i_{n_i})}\}$, $D_i \cap D_j = \phi, (i \neq j)$, $n_1 + n_2 + \cdots + n_m = n'$.

In general, the amount of observations $\{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\}$ is huge and these observations are not used in total to determine the knowledge fluctuation zone and the correctness factor. Rather, the sampling techniques may be applied. More specifically, we may follow the following the procedure to do parepeatic software data mining.

Step (1). Obtain the dataset $\{X^{(1)}, X^{(2)}, \ldots, X^{(n)}\}$ from n elements of the given parepeatic universe $U^{(p)}$, where $X^{(j)} = [x_1^{(j)}, x_2^{(j)}, \ldots, x_5^{(j)}]^\tau$

Step (2). Remove all the data points which satisfy the condition $x_1 x_2 \cdots x_5 = 0$, that is, the data points with no usages of any operators or operands are treated as outliers. The resulting dataset with outliers being removed is denoted as $\{T^{(1)}, T^{(2)}, \ldots, T^{(n)}\}$, with $T^{(j)} = [t_1^{(j)}, t_2^{(j)}, \ldots, t_5^{(j)}]^\tau$.

Step (3). Determine the maximal number of disjoint classes that the dataset is allowed to be divided. Denote this number as N.

Step (4). Determine the number of sampling sets which are to be obtained from $\{T^{(1)}, T^{(2)}, \ldots, T^{(n)}\}$; let the number be κ.

Step (5). Let $\alpha = 1$.

Step (6). Generate a random positive integer over the range $[s_1, s_2]$; this number is denoted as $s^{(\alpha)}$.

Step (7). Sample $s^{(\alpha)}$ data points one by one from the dataset $\{T^{(1)}, T^{(2)}, \ldots, T^{(n)}\}$ without replacement; that is, if $T^{(1)}$ is sampled, then it will not be returned back to the sampled dataset and thus the resulting $s^{(\alpha)}$ data points must be distinct; let the resulting $s^{(\alpha)}$ data points make up the dataset $\{\Gamma^{(1)}, \Gamma^{(2)}, \ldots, \Gamma^{(s^{(\alpha)})}\}$, with $T^{(j)} = [\gamma_1^{(j)}, \gamma_2^{(j)}, \ldots, \gamma_5^{(j)}]^\tau$.

Step (8). Transform the dataset $\{\Gamma^{(1)}, \Gamma^{(2)}, \ldots, \Gamma^{(s^{(\alpha)})}\}$ into $\{L^{(1)}, L^{(2)}, \ldots, L^{(s^{(\alpha)})}\}$ such that $L^{(i)} = \gamma_2^{(i)} \log_2 \gamma_2^{(i)} + \gamma_3^{(i)} \log_2 \gamma_3^{(i)}$, $i = 1, 2, \ldots, s^{(\alpha)}$; that is, $L^{(i)}$ is the Halstead length of a C language program.

Step (9). Transform the dataset $\{L^{(1)}, L^{(2)}, \ldots, L^{(s^{(\alpha)})}\}$ into $\{Y^{(1)}, Y^{(2)}, \ldots, Y^{(s^{(\alpha)})}\}$ such that $Y^{(i)} = \dfrac{L^{(i)} - \min\limits_{1 \le i \le \alpha} L^{(i)}}{\max\limits_{1 \le i \le \alpha} L^{(i)} - \min\limits_{1 \le i \le \alpha} L^{(i)}}$, $i = 1, 2, \ldots, s^{(\alpha)}$; that is, the dataset is normalized and $Y^{(i)} \in [0,1]$, $i = 1, 2, \ldots, s^{(\alpha)}$.

Step (10). Cluster the dataset $\{Y^{(1)}, Y^{(2)}, \ldots, Y^{(s^{(\alpha)})}\}$ into a number of disjoint classes, say, $(C_1^{(\alpha)}, C_2^{(\alpha)}, \ldots, C_{n_\alpha}^{(\alpha)})$ according to some clustering algorithm (e.g., fuzzy clustering algorithm); there hold $C_j^{(\alpha)} \cap C_k^{(\alpha)}$ for $j \neq k$, $C_j^{(\alpha)} = \{r_{j1}^{(\alpha)}, r_{j1}^{(\alpha)}, \ldots, r_{jd_j}^{(\alpha)}\}$, and $C_1^{(\alpha)} \cup C_2^{(\alpha)} \cup \cdots \cup C_{n_\alpha}^{(\alpha)} = \{Y^{(1)}, Y^{(2)}, \ldots, Y^{(s^{(\alpha)})}\}$.

Step (11). Evaluate the performance of the resulting clustering $(C_1^{(\alpha)}, C_2^{(\alpha)}, \ldots, C_{n_\alpha}^{(\alpha)})$ as

$$p_\alpha = \begin{cases} w_1 R^{(\alpha)} + \dfrac{w_2}{D^{(\alpha)}} + \dfrac{w_3}{N} & \text{if} \quad n_\alpha \le N \\ \infty & \text{otherwise} \end{cases}$$

where $w_1, w_2, w_3 \in [0,1]$ are weighting coefficients, with $w_1 + w_2 + w_3 = 1$, $R^{(\alpha)} = \max\limits_{1 \le j \le n_\alpha} \max\limits_{1 \le \beta, \nu \le d_j} \left| r_{j\beta}^{(\alpha)} - r_{j\nu}^{(\alpha)} \right|$, $D^{(\alpha)} = \min\limits_{1 \le i, j \le n_\alpha} \min\limits_{1 \le \beta \le d_i, 1 \le \nu \le d_j} \left| r_{i\beta}^{(\alpha)} - r_{j\nu}^{(\alpha)} \right|$; that is, $R^{(\alpha)}$ is the maximum of the diameter of a class, and $D^{(\alpha)}$ the minimum of the distances between two distinct classes. A good clustering should lead to small $R^{(\alpha)}$ and large $D^{(\alpha)}$. So, the smaller $p^{(\alpha)}$, the better the clustering.

Step (12). Obtain the result of the clustering as a data pair (n_α, p_α).

Step (13). Let $\alpha = \alpha + 1$; if $\alpha \le \kappa$, go to Step (6).

Step (14). Obtain the result of data sampling and processing as a new dataset $((n_1, p_1), (n_2, p_2), \ldots, (n_\kappa, p_\kappa))$.

Step (15). Obtain the knowledge fluctuation zone by removing the possible outliers from the dataset $((n_1, p_1), (n_2, p_2), \ldots, (n_\kappa, p_\kappa))$; $E(X) = ((n_{1'}, p_{1'}), (n_{2'}, p_{2'}), \ldots, (n_{\kappa'}, p_{\kappa'}))$, with $\kappa' \le \kappa$; that is, the dataset $((n_1, p_1), (n_2, p_2), \ldots, (n_\kappa, p_\kappa))$ is divided into two classes, the fluctuation zone and the outliers, by using some clustering algorithm (e.g., fuzzy clustering algorithm).

Step (16). Transform the dataset $((n_{1'}, p_{1'}), (n_{2'}, p_{2'}), \ldots, (n_{\kappa'}, p_{\kappa'}))$ into $((\eta_{1'}, p_{1'}), (\eta_{2'}, p_{2'}), \ldots, (\eta_{\kappa'}, p_{\kappa'}))$, with $\eta_\alpha = \dfrac{n_\alpha}{N}$.

Step (17). Let $\overline{\eta} = \frac{1}{\kappa'}\sum_{\alpha=1}^{\kappa'}\eta_\alpha$, $\overline{p} = \frac{1}{\kappa'}\sum_{\alpha=1}^{\kappa'} p_\alpha$

Step (18). Let $\varsigma = \arg\min_{1\le\alpha\le\kappa'}\sqrt{(\eta_\alpha-\overline{\eta})^2+(p_\alpha-\overline{p})^2}$; that is, $(\eta_\varsigma, p_\varsigma)$ is the one that is closest to $(\overline{\eta},\overline{p})$.

Step (19). Choose $(\eta_\varsigma, p_\varsigma)$ as the central knowledge, that is, $f(X)=(\eta_\varsigma, p_\varsigma)$.

Step (20). Determine the number of sampling sets which are to be obtained from $\{T^{(1)}, T^{(2)}, \ldots, T^{(n)}\}$; let the number be κ_c.

Step (21). Re-perform Steps (5) to (13) except that κ is replaced by κ_c.

Step (22). Obtain the output of Step (21) as a new dataset $((l_1,q_1),(l_2,q_2),\ldots,(l_{\kappa_c},q_{\kappa_c}))$ as that obtained in Step (14); this dataset is used to assess the knowledge fluctuation zone and obtain the corresponding correctness factor.

Step (23). Transform the dataset $((l_1,q_1),(l_2,q_2),\ldots,(l_{\kappa_c},q_{\kappa_c}))$ into $((\omega_1,q_1),(\omega_2,q_2),\ldots,(\omega_{\kappa_c},q_{\kappa_c}))$, with $\omega_\alpha = \frac{l_\alpha}{N}$.

Step (24). Let $\varepsilon_i = \min_{1\le j\le\kappa'}\sqrt{(\omega_i-\eta_j)^2+(q_i-p_j)^2}$, where (η_j, p_j) is specified in Step (16); ε_i measures the distance between the data point (l_i, q_i) and the knowledge fluctuation zone.

Step (25). Determine a distance threshold ε_0.

Step (26). Determine the correctness factor as $c = \frac{1}{\kappa_c}\sum_{i=1}^{\kappa_c}\delta_i$, where

$$\delta_i = \begin{cases} 1 & \text{if } \varepsilon_i \le \varepsilon_0 \\ 0 & \text{otherwise} \end{cases}$$

Step (27).

End.

Remarks

(1). The central knowledge we want to extract from the given dataset by using the above algorithm is an approximate clustering. Different intended central knowledge should lead to different data mining algorithm. In our previous work [7], the intended central knowledge is a functional relation among several feature variables.

(2). Data pre-processing takes place in Step (2). However it also takes place in Steps (7), (8), (9) and (21). That is, besides in the beginning of data mining, data pre-processing may take place throughout the rest process of data mining.

(3). Steps (3) to (15) are devoted to determining the fluctuation zone, Steps (16) to (19) to determining the central knowledge, and Steps (20) to (26) to determining the correctness factor.

(4). The sampling technique plays a major role in the parepeatic data mining algorithm given above. It is used both in determining the fluctuation zone and in determining the correctness factor. The dataset $((n_1,p_1),(n_2,p_2),\ldots,(n_\kappa,p_\kappa))$ obtained in Step (14) can be treated as the training dataset for the parepeatic model of the intended knowledge, whereas the dataset $((l_1,q_1),(l_2,q_2),\ldots,(l_{\kappa_c},q_{\kappa_c}))$ obtained in Step (22) can be treated as the validation dataset for the parepeatic model.

(5). The performance evaluation criterion p_α adopted in Step (11) is key part of the intended knowledge. It evaluates how good the clustering is. It makes trade-offs among the diameters of a class, the distances among distinct classes, and the number of distinct classes by using the weighting coefficients w_1, w_2 and w_3. The allowed maximal number of distinct classes, N , is also required for the performance evaluation.

(6). In determining the correctness factor c, the distance threshold ε_0 must be specified. Different values of ε_0 will lead to different values for c.

(7). Besides the parameters $w_1, w_2, w_3, N, \varepsilon_0$, the algorithm given above also requires the positive integer interval $[s_1, s_2]$ to be specified for the sampling processes of determining the fluctuation zone and the correctness factor.

REFERENCES

[1] U. Fayyad, P. Stolotz. "Data Mining and KDD: Promise and Challenges", *Future Generation Computer Systems*, Vol.13, 1997, pp99-115.

[2] J. Han, "Data Mining", in: J.Urban, P.Dasgupta (eds), *Encyclopedia of Distributed Computing*, Kluwer Academic Publishers, 1999.

[3] X. Wu, "Building Intelligent Learning Database Systems", *AI Magazine*, Vol.21, No.3, 2000, pp59-65.

[4] C. Zhang, S. Zhang, *Association Rule Mining: Models and Algorithms*, Springer, 2002.

[5] A. Jain, M. Murty, P. Flynn, Data Clustering: a Review", *ACM Computing Survey*, Vol.31, No.3, 1999, pp264-323.

[6] K. Y. Cai, J. H. Liao, "Software Pattern Laws and Partial Repeatability" in: G. Q. Chen, M. S. Ying, K. Y. Cai, (editors.) *Fuzzy Logic and Soft Computing*, Kluwer Academic Publishers, 1999, pp89-120.

[7] K. Y. Cai, L. Chen, "Analyzing Software Science Data with Partial Repeatability", *Journal of Systems and Software*, Vol.63, 2002, pp173-186.

[8] J. Granmeier, A. Rudolph, "Techniques of Cluster Algorithms in Data Mining", *Data Mining and Knowledge Discovery*, Vol.6, 2002, pp303-360.

[9] Y. Li, N. Zhong, "Interpretations of Association Rules by Granular Computing," *Proc. the Third IEEE International Conference on Data Mining*, 2003, pp.593-596.

[10] M. H. Halstead, *Elements of Software Science*, Elsevier, 1977.

Web-based Multi-Criteria Group Decision Support System with Linguistic Term Processing Function

Jie Lu, Guangquan Zhang & Fengjie Wu

Abstract— **Organizational decisions are often made in groups where group members may be distributed geographically in different locations. Furthermore, a decision-making process, in practice, frequently involves various uncertain factors including linguistic expressions of decision makers' preferences and opinions. This study first proposes a rational-political group decision-making model which identifies three uncertain factors involved in a group decision-making process: decision makers' roles in a group reaching a satisfactory solution, preferences for alternatives and judgments for assessment-criteria. Based on the model, a linguistic term oriented multi-criteria group decision-making method is developed. The method uses general fuzzy number to deal with the three uncertain factors described by linguistic terms and aggregates these factors into a group satisfactory decision that is in a most acceptable degree of the group. Moreover, this study implements the method by developing a web-based group decision support system. This system allows decision makers to participate a group decision-making through the web, and manages the group decision-making process as a whole, from criteria generation, alternative evaluation, opinions interaction to decision aggregation. Finally, an application of the system is presented to illustrate the web-based group decision support system.**

Index Terms— **Decision support systems, Group decision-making, Fuzzy decision-making, Web-based systems, Linguistic terms**

I. Introduction

MANY organizational decisions are made through evaluating a set of alternatives and then selecting the most satisfactory one from them based on the information at hand and the perspectives of decision makers. These alternatives may exist objectively such as a number of candidates for a position, or is nominated by decision makers such as several proposals for new product development, or is generated using a suitable decision model such as multi-objective programming. Multiple criteria are often used to evaluate the set of alternatives where some criteria could be more important than others in selecting the most satisfactory one. Also, in organizations, many decisions and the processes involved in making them are performed at a group level rather an individual, refereed to group decision-making (GDM) [1]. A group decision-making process is to find a group satisfactory solution which is one that is most acceptable by the group of individuals as a whole.

Three basic factors may influence the assessment of utility of alternatives and the deriving of the group satisfactory solution. The first one is individual's role (weight) in the selection of the satisfactory solution. The second factor is individual's preference for alternatives. The third factor is criteria used for assessing these alternatives and judged by decision makers [2]. The three factors are often expressed by linguistic terms in a group decision-making practice. For example, an individual role can be described using linguistic terms 'important person' or 'general person'. Similar, to express a decision maker's preference for an alternative linguistic terms 'low' and 'high' could be used, and to express a decision maker's judgment for comparison of a pair of assessment-criteria 'equally important' or 'A is more important than B' are often applied. As linguistic terms are too complex and ill-defined to be reasonably described in conventional quantitative expressions [3], a crucial requirement is proposed for linguistic information processing technique. The concept of linguistic variable was proposed by Zadeh [4] to deal with the situations and is described and operated by fuzzy set theory.

Due to the computational complexity of GDM methods, decision support systems (DSS) have been applied as a support tool for solving GDM problems [5], referred to group decision support systems (GDSS). When a GDSS applies fuzzy set technology to handle uncertainty issues it is normally referred to fuzzy GDSS (FGDSS) [6]. Traditionally, DSS, including GDSS, had to be installed in a specified location, such as a decision room. Now, the web is acting as a mechanism for the support of decision-making in organizations, particularly geographically distributed organizations [7, 8]. GDSS can therefore be implemented as a kind of web-based services, and thus have been moving to a global environment. Since the advance of web technology, which allows users fast and inexpensive access to an unprecedented amount of information provided by websites, digital libraries and other data sources, web-based DSS have been applied in a widespread decision activities with its unified graphical user interface [7, 9]. Although existing literature provides a way to build web-based GDSS, such as Wang and Chien [8], there is no report regarding to build a web-based GDSS to deal with decision members' linguistic term enter and processing.

This paper first establishes a rational-political group decision-making model which identifies three uncertain factors involved in a group decision-making process. It then proposes a linguistic term based group decision-making method to handle wholly the three fuzzy properties (uncertainty in decision makers' roles for reaching a satisfactory solution, their

Authors are with Faculty of Information Technology, University of Technology, Sydney, PO Box 123, Broadway, NSW 2007, Australia (e-mail: {jielu, zhangg, fengjiew}@it.uts.edu.au).

preferences for alternatives and their judgments for assessment-criteria) simultaneously in a group decision-making. It uses general type of fuzzy numbers to describe linguistic terms, so that users can choose any type of fuzzy number in applications. It also applies inference rules correcting inconsistency in individual preference explanation. Based on the method, a web-based fuzzy group decision support system (WFGDSS) is developed. An initial experiment shows that the WFGDSS can improve the effectiveness and application range of group decision-making, and use of linguistic terms can increase users' confidence in deriving a satisfactory solution from a set of alternatives in a group.

The rest of this paper is organised as follows. Section 2 shows a rational-political model of group decision-making with uncertainty and analyzes the features of web-based GDSS. Section 3 gives a fuzzy group decision-making method. The WFGDSS and an application for using the system are shown in Section 4. Conclusions are given in Section 5.

II. RATIONAL-POLITICAL GROUP DECISION-MAKING MODEL WITH UNCERTAINTY AND WEB FEATURES

Through literature reviewing this section proposes a rational-political group decision support model with uncertainty and analyses the main features of web-based decision support systems.

A. Rational-political group decision support model with uncertainty

Group decision-making is a key component to the functioning of an organization, because organizational performance involves more than just individual actions. It is the process of arriving at a satisfactory solution based upon the input and feedback of multiple individuals. It, therefore, is very important to determine what makes group decision-making effective and to increase the level of overall satisfaction for the final decision across the group [1]. Due to the importance and complexity of the group decision-making process, decision-making models are needed to establish a systematic means of supporting effective and efficient group decision-making [10].

There are two kinds of most popular and basic models of group decision-making. The first one is the rational model [11]. The kind of models is grounded on objectives, alternatives, consequences and optimality. It assumes that complete information regarding the decision to be made is available and one correct conception of the decision can be determined. Another kind of decision-making models is the political model. In contrast to the rational model, the individuals involved do not accomplish the decision task through rational choice in regard to business objectives. The decision makers are motivated by and act on their own needs and perceptions. This process involves a cycle of negotiation and idea sharing among the group members in order, for each one, to try to get his or her perspective to be the one of choices. More specifically, this process involves each decision maker trying to sway powerful people (such as a group leader) within the situation to adopt his or her viewpoint and influence the remaining members [11, 12].

In a real group decision-making process of an organization, decision makers are often involved in a group discussion to express their opinions for convincing other members and influencing final group decision. Obviously, decision makers' opinions will directly impact on the assessment of utility of alternatives and the deriving of an optimal group decision. In such a situation, the group optimal decision is in reality the group satisfactory decision. Three main factors regarding to decision maker opinion have been identified with a direct influence for the form of an optimal group decision [2].

The first one is individual's role (weight) in the selection of the optimal decision. There may be a group leader or leaders who play more important roles than others in a particular group decision-making. Although each decision maker tries to influence other members to adopt his or her viewpoint, powerful members will sway strongly the decision-making than other members. Group members thus have different 'weights' in a group decision-making, and the situation should be reflected on the generation process of the group satisfactory decision.

The second factor is individual's preference for alternatives. Group members may not know all information related to a decision problem or may not consider all relevant information to the decision problem. Also, they may have different understanding for same information, different experience in the area of current decision problem, and, therefore, different preferences for alternatives. The different preferences of group members impact directly on the deriving of the group optimal decision.

The third factor is criteria for assessing these alternatives. Assessment-criteria are usually determined through generation and discussion in decision groups. Goals or priorities of decision objectives are often as assessment-criteria for multi-objective decision problems. In a real situation, different group members may have different viewpoints in assessment-criteria for a decision problem because of workload, time and inexperience at assessing a problem all affect determining assessment-criteria. Different members may often have different judgments in comparing the importance between a pair of assessment-criteria. Obviously, what assessment-criteria are used and how priority of each assessment-criterion is will directly influence the selection of the group's satisfactory decision.

To deal with the three factors and support the achievement of consensus of group decision-making in a real environment, this research proposes a rational-political model which combines the advantages of both rational and political models. By inheriting the optimization property of rational model, it shows a sequential approach to make a group decision. By carry out the advantages of political model, it allows decision makers to have inconsistent assessment, incomplete information and inaccurate opinions for alternatives and assessment-criteria. The model, therefore, can deal with the three uncertain factors simultaneously.

As shown in Figure 1, the model is assumed that a set of alternatives for a decision problem has been conducted. A number of group members, including a group leader, will work together to select an optimal solution from these alternatives. A set of assessment-criteria for assessing these alternatives are nominated by these group members or generated through running a suitable model operated by them. Group members (including the leader) are awarded or assigned weights before or at the beginning of the decision-making process. It is often done by the leader. Although group members may have different experiences, opinions and information at hand for the decision problem, they must participate in the group aggregating process to ensure that the disparate individuals come to share the same decision objectives. These group members will be required to give their individual judgments for priority of proposed assessment-criteria and preferences for alternatives under these assessment-criteria by linguistic terms. The final group decision is made through optimizing and aggregating group members' preferences on alternatives under their weights and judgments on assessment-criteria.

B. Fuzzy group decision-making methods

The aggregation of group members' perceptions involves the presentation and operation of linguistic terms. Zadeh's fuzzy set theory [4] is naturally applied in the aggregation process with uncertainty and imprecision. Several typical fuzzy group decision-making methods have been developed and focused respectively on the three uncertain factors. Some researches such as [3], [13], [14], [15], and [16] have been done in describing the uncertainty of individual preferences for alternatives and aggregating imprecise individual preferences into a group consensus decision. The uncertainty on the judgment of assessment-criteria has also been paid attentions by researchers such as [17], [18] and [19]. The uncertainty of individual roles, or call it individual weights, in attempting to reach a group satisfactory solution has been discussed in the literature of this area such as [20] and [21]. Furthermore, our earlier research [2] has proposed a framework to identify uncertainty and imprecise related to the three factors and find a possible way to provide a representation of group members' perspectives in order to minimize their conflict in a decision-making process.

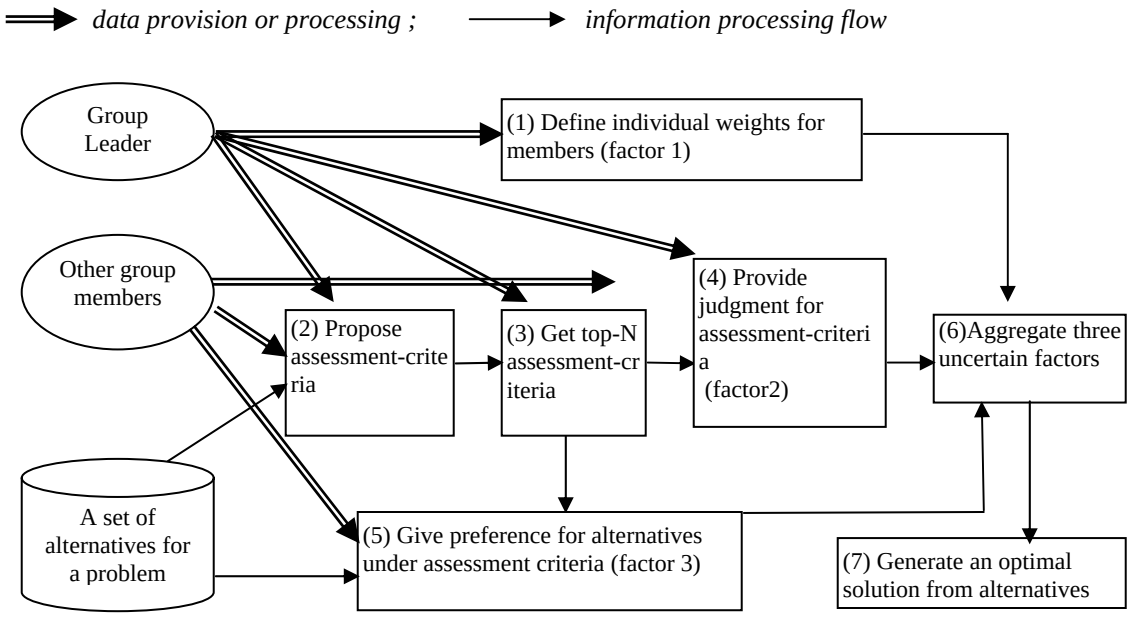


Fig. 1: Rational-political group decision-making model with uncertain factor

C. Features of web-based group decision support systems

GDSS, as a 'specific' type of systems within the broad family of DSS [22], have been successfully implemented in many organizations at different organizational levels [23]. Group members may be distributed in different locations therefore a group decision meeting need web-based tool support. Four features of web-based group decision support systems are identified as follows.

Supporting asynchronous communication among group members: An important feature provided by GDSS is to support interpersonal communication and coordination among group members. This feature aims at achieving a common understanding of the issues revealed and arriving at a group satisfactory decision. The communication and coordination activities of group members are facilitated by technologies that can be characterized along the three continua of time, space, and level of group support [24]. In general, group members could only communicate synchronously by face-to-face meetings without web technology. With the applications of the web and specific web-based GDSS, group members can

communicate asynchronously in group decision-making, and also can obtained information through emails, bulletin board systems, Internet newsgroups and other web applications.

Extending application range of GDSS: Web-based GDSS can use web environment as a development and delivery platform [22]. More recently, both e-business and e-government are increasing their demands for more online data analysis and decision support. The web platform, which is also a platform for e-business and e-government development, lends web-based GDSS to have widespread use and adoption in organizations. Also, organizations can use web-based GDSS proving group decision support capability to managers over a proprietary Intranet, to customers and suppliers over an Extranet, or to any stakeholder over the global Internet.

Reducing technological barriers: web-based GDSS can reduce technological barriers and make less costly to develop and delivery itself and provide decision-relevant information [24]. Traditionally, GDSS required specific software on user computers, specific locations to set up, and users needed proper training to learn how to use a GDSS. From the web platform, GDSS do not require any specific support in software, location and user training. Further, by using the web, GDSS have a convenient and graphical user interface with visualization possibilities, and therefore are automatically available to large number of decision makers. As a result, managers who have not used GDSS before will find web-based GDSS powerful and convenient. Managers who have been exposed to traditional GDSS tools in the 1980s and 1990s will find that web-based GDSS have provided more support that the traditional techniques could not, including easily accessible and unique user interface.

Improving effectiveness of decision-making performance: Building web-based GDSS can increase the range and depth of information access, and therefore improve the solving of decision problems and the effectiveness of decision-making performance [25]. Decision-making, especially at upper management levels, relies heavily on data sources outside the organization. The web-based GDSS by using web mining and related web intelligence techniques allow decision makers to access internal and external data sources, such as competitor's product/service offerings, during the decision-making process. In particular, the organizations will find that web-based GDSS can more effectively assist their decision groups in making organizational strategic decisions where group members are distributed in different locations [22].

There is sufficient evidence showing that web-based GDSS can extend the applications of traditional GDSS and support more effectively organizational decision-making performance [26]. A number of web-based GDSS have been developed in the last few years. These include GEO-ELCA which is a web-based collaborative spatial DSS [27] and an agent-based Internet-based GDSS [8]. The development of our WFGDSS will extend current results by proving the ability of dealing with linguistic terms using general fuzzy number technique.

III. Linguistic Term Oriented Fuzzy Multi-Criteria Group Decision-Making Method

This section introduces a fuzzy multi-criteria group decision-making method which consists of night steps within three levels.

Let $\boldsymbol{S} = \{\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_m\}$, $m \geq 2$, be a given finite set of alternative solutions for a decision problem, and $\boldsymbol{P} = \{\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_n\}$, $n \geq 2$, be a given finite set of group members to select a satisfactory solution from $\boldsymbol{S}$. The proposed method is described as follows.

Level 1: *Assessment-criteria and individual weight generation*

Step 1: Each group member $\boldsymbol{P}_k$ (k = 1, 2, ..., n) can propose one or more assessment-criteria $(C^k_{k_1}, C^k_{k_2}, \cdots, C^k_{k_p})$, $p=1, 2\ldots w$, for selecting a solution from alternatives. All members' assessment-criteria are put into a criterion pool and top-T criteria, $\boldsymbol{C} = \{\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_t\}$, are chosen as assessment-criteria for the decision problem in the group.

Step 2: As group members play different roles in an organization and therefore have different degree of influence for the selection of the satisfactory solution. That means the relative importance of each group member may not equal in a decision group. Some members, in particular the group leader, are more powerful than the others for a specific decision problem. Therefore, in the method, each member is assigned with a weighting that is described by a linguistic term $\tilde{v}_k, k = 1, 2, \cdots, n$. These terms are determined through discussion in the group or assigned by higher management level before or at the beginning of the decision process. For example, $\boldsymbol{P}_k$ is assigned with 'strongly important person (SP)'. Possible linguistic terms used in the factor are shown in Table 1.

TABLE 1.
LINGUISTIC TERMS USED FOR DESCRIBING WEIGHTS OF GROUP MEMBERS

Linguistic terms	Fuzzy numbers
General decision person (GP)	f_1
Weakly important person (WP)	f_2
Strongly important person (SP)	f_3
The most important person (TP)	f_4

Level 2: *Individual Preference and Judgment Generation*

Step 3: Each group member $\boldsymbol{P}_k$ (k = 1, 2, ..., n) is required to express his/her opinion for assessment-criteria by pairwise comparison of the relative importance of these criteria using Analytic Hierarchy Process (AHP) method.

An initial pairwise comparison matrix $\boldsymbol{E} = [\tilde{e}^k_{ij}]_{t\times t}$ is first established, where $\tilde{e}^k_{ij}$ represents the quantified judgments on pairs of assessment-criteria $\boldsymbol{C}_i$ and $\boldsymbol{C}_j$ (i, j =1, 2, …, t, $i \neq j$). The comparison scale belongs to a set of linguistic terms that contain various degrees of preferences required by the group member $\boldsymbol{P}_k$ (k = 1, 2, ..., n), or take a value '*'. The linguistic terms are show in Table 2. Character '*' represents that group member $\boldsymbol{P}_k$ (k = 1, 2, ..., n) doesn't know or cannot compare the

relative importance of assessment-criteria $\boldsymbol{C}_i$ and $\boldsymbol{C}_j$.

TABLE 2.
LINGUISTIC TERMS USED FOR THE COMPARISON OF ASSESSMENT-CRITERIA

Linguistic terms	Fuzzy numbers
Absolutely more unimportant (ANI)	a_1
Strongly more unimportant (SNI)	a_2
Weakly more unimportant (WNI)	a_3
Equally important (EI)	a_4
Weakly more important (WI)	a_5
Strongly more important (SI)	a_6
Absolutely more important (AI)	a_7

By using following linguistic variable inference rules, the inconsistence of each pairwise comparison matrix $\boldsymbol{E} = [\tilde{e}_{ij}^{\,k}]_{t\times t}$ is corrected:

Positive-Transitive rule: If $\tilde{e}_{ij}^{\,k} = a_s$ (s = 4, 5, 6, 7), and $\tilde{e}_{jm}^{\,k} = a_t$ (t = 4, 5, 6, 7), then $\tilde{e}_{im}^{\,k} = a_{\max(s,t)}$. For example, if $\boldsymbol{C}_i$ is '*equally important*' with $\boldsymbol{C}_j$ (s = 4), and $\boldsymbol{C}_j$ is '*strongly more important*' with $\boldsymbol{C}_m$ (t = 6) then $\boldsymbol{C}_i$ is '*strongly more important*' with $\boldsymbol{C}_m$.

Negative-Transitive rule: If $\tilde{e}_{ij}^{\,k} = a_s$ (s = 3, 2, 1), and $\tilde{e}_{jm}^{\,k} = a_t$ (t = 3, 2, 1), then $\tilde{e}_{im}^{\,k} = a_{\min(s,t)}$. For example, $\boldsymbol{C}_i$ is '*absolutely more unimportant*' than $\boldsymbol{C}_j$ (s = 1), $\boldsymbol{C}_j$ is a '*weakly more unimportant*' than $\boldsymbol{C}_m$ (t = 3), then $\boldsymbol{C}_i$ is '*absolutely more unimportant*' than $\boldsymbol{C}_m$.

De-In-Uncertainty rule: If $\tilde{e}_{ij}^{\,k} = a_s$ (s = 4, 5, 6, 7), $\tilde{e}_{jm}^{\,k} = a_t$ (t = 3, 2, 1,) or *, then $\tilde{e}_{im}^{\,k} = a_i$ for any $t \le i \le s$ or *. For example, $\boldsymbol{C}_i$ is '*weakly more important*' with $\boldsymbol{C}_j$ (s = 5) and $\boldsymbol{C}_j$ is '*strongly more unimportant*' with $\boldsymbol{C}_m$ (t = 2), then $\boldsymbol{C}_i$ can have any relationship between '*strongly more unimportant*' and '*weakly more important*', such as '*equally important* (i = 4)' or '*', with $\boldsymbol{C}_m$.

In-De-Uncertainty rule: If $\tilde{e}_{ij}^{\,k} = a_s$, ($s$ = 3, 2, 1) or *, and $\tilde{e}_{jm}^{\,k} = a_t$ (t = 4, 5, 6, 7), then $\tilde{e}_{im}^{\,k} = a_i$ for any $s \le i \le t$, or *. For example, $\boldsymbol{C}_i$ is '*weakly more unimportant*' with $\boldsymbol{C}_j$ (s = 3) and $\boldsymbol{C}_j$ is '*strongly more important*' with $\boldsymbol{C}_m$ (t = 6) then $\boldsymbol{C}_i$ can have any relationship between '*weakly more unimportant*' and '*strongly more important*', such as '*equally important* (i = 4)' or '*', with $\boldsymbol{C}_m$,

Consistent weights w_i^k (i = 1, 2, ⋯, t) for every assessment-criterion can be determined by calculating the geometric mean of each row of the matrix $[\tilde{e}_{ij}^{\,k}]_{t\times t}$ where e_{ij}^k (j = 1, 2, ⋯, i_k) is not '*', and then the resulting fuzzy numbers are normalized and denoted as $\tilde{w}_1^k, \tilde{w}_2^k, \cdots, \tilde{w}_t^k$, where $\tilde{w}_i^k \in F_T^*(R)$ and

$$\tilde{w}_i^k = \frac{w_i^k}{\sum_{i=1}^{t} {w_i^k}^R_0}, \quad \text{for } i = 1, 2, \cdots, t; k = 1, 2, \cdots, n. \tag{10}$$

Step 4: Against every assessment-criterion $\boldsymbol{C}_j$ (j = 1, 2, ..., t), a belief level can be introduced to express the possibility of selecting a solution $\boldsymbol{S}_i$ under criterion $\boldsymbol{C}_j$ for a group member $\boldsymbol{P}_k$. The belief level b_{ij}^k (i = 1, 2, ⋯, t, j = 1, 2, ⋯, m, k = 1, 2, ..., n) belongs to a set of linguistic terms that contain various degrees of preferences required by a group member $\boldsymbol{P}_k$ (k = 1, 2, ..., n) under jth assessment-criterion (j = 1, 2, ..., t). The linguistic terms for variable 'preference' are shown in Table 3. Notation '**' can be used to represent that group member $\boldsymbol{P}_k$ doesn't know or could not give a belief level for expressing the preference for a solution $\boldsymbol{S}_i$ under assessment-criterion $\boldsymbol{C}_j$.

Step 5: Belief level matrix $(b_{ij}^k)_{(k=1,2,\cdots,n)}$ is aggregated in to belief vector $(\bar{b}_j^{\,k})$ (j = 1, 2, ..., m, k = 1, 2, ..., n).

$$\bar{b}_j^{\,k} = \tilde{w}_{j_1}^k * b_{jj_1}^k + \tilde{w}_{j_2}^k * b_{jj_2}^k + \cdots + \tilde{w}_{j_s}^k * b_{jj_s}^k, \tag{11}$$

where $b_{jj_i}^k$ (i = 1, 2, ⋯, s) is not '**'. Based on belief vectors $(\bar{b}_j^{\,k})$, the group member $\boldsymbol{P}_k$ (k = 1, 2, ..., n) can make an overall judgment on the alternatives, called an individual assessment vector. All individual selection vectors can compose a group of selection matrixes $(\bar{b}_j^{\,k})_{n\times m}$.

TABLE 3.
LINGUISTIC TERMS USED FOR PREFERENCE BELIEF LEVELS FOR ALTERNATIVES

Linguistic terms	Fuzzy numbers
Very low (VL)	b_1
Low (L)	b_2
Medium low (ML)	b_3
Medium (M)	b_4
Medium high (MH)	b_5
High (H)	b_6
Very high (VH)	b_7

Level 3: *Group Decision Aggregation*:

Step 6: As each member $\boldsymbol{P}_k$ has been assigned with a weighting $\tilde{v}_k$, $k = 1, 2, \cdots, n$ as shown in Table 1, a weight vector is obtained:

$$V = \{\tilde{v}_k,\ k = 1, 2, \cdots, n\}.$$

The normalized weight of a group member $\boldsymbol{P}_k$ (k = 1, 2, ..., n) is denoted as

$$\tilde{v}_k^* = \frac{\tilde{v}_k}{\sum_{i=1}^{n} {v_i}^R_0}, \quad \text{for } k = 1, 2, \cdots, n. \tag{12}$$

Step 7: Considering the normalized weights of all group members, we can construct a weighted normalized fuzzy decision vector

$$(\tilde{r}_1, \tilde{r}_2, \cdots, \tilde{r}_m) = (\tilde{v}_1^*, \tilde{v}_2^*, \cdots, \tilde{v}_n^*) \begin{pmatrix} \bar{b}_1^1 & \bar{b}_2^1 & \cdots & \bar{b}_m^1 \\ \bar{b}_1^2 & \bar{b}_2^2 & \cdots & \bar{b}_m^2 \\ \vdots & \vdots & \ddots & \vdots \\ \bar{b}_1^n & \bar{b}_2^n & \cdots & \bar{b}_m^n \end{pmatrix}, \tag{13}$$

where $\tilde{r}_j = \sum_{k=1}^{n} \tilde{v}_k^* \bar{b}_j^k$.

Step 8: In the weighted normalized fuzzy decision vector the elements $\tilde{v}_j$, j = 1,2, ⋯, m , are normalized positive fuzzy numbers and their ranges belong to closed interval [0, 1]. We can then define fuzzy positive-ideal solution (FPIS, r^*) and

fuzzy negative-ideal solution (FNIS, r^-) as:

$$r^* = 1 \quad \text{and} \quad r^- = 0.$$

The positive and negative solution distances between each $\tilde{r}_j$ and r^*, $\tilde{r}_j$ and r^- can be calculated as:

$$d_j^* = d(\tilde{r}_j, r^*) \quad \text{and} \quad d_j^- = d(\tilde{r}_j, r^-), \quad j = 1, 2, \cdots, m, \qquad (14)$$

where $d(.,.)$ is the distance measurement between two fuzzy numbers.

Step 9: A closeness coefficient is defined to determine the ranking order of all solutions once the d_j^* and d_j^- of each decision solution $\boldsymbol{S}_j$ (j = 1, 2, ..., m) are obtained. The closeness coefficient of each solution is calculated as:

$$CC_j = \frac{1}{2}\left(d_j^- + (1 - d_j^*)\right), \quad j = 1, 2, \cdots, m. \qquad (15)$$

The solution $\boldsymbol{S}_j$ that corresponds to the $Max(CC_j$, j=1, 2, …, $m)$ is the satisfactory solution of the decision group.

If the selected solution cannot be accepted by the decision group two actions can be taken. One is to change assessment-criteria particularly when further information is available, and another is to remove the worst alternative solution and redo the decision-making process. The 'worst' solution is one that corresponds to the $Min(CC_j;\ j = 1, 2, \ldots, m)$.

IV. WFGDSS AND ITS APPLICATION

This section presents the design and implementation of WFGDSS. An illustrated example is given to demonstrate the application of WFGDSS.

A. Architecture and working process of the WFGDSS

The architecture of WFGDSS is shown in Fig. 2. The web server manages all web pages of the system, traces user information and provides simultaneously services to multiple group members through sessions, applications and coking facilities. All web pages developed in WFGDSS, for interacting dynamically to group members in solving multi-criteria group decision-making problems with linguistic terms, are created on the fly by the web server. Using a server side application program, the web server can manage and implement client tasks. The database sever interacts with the web sever by using an ODBC connection. The system is developed and implemented mainly in JSP combined with HTML and JavaScript.

The working process of a decision group using WFGDSS is described as follows.

The group leader first uses a browser to log in the system and define a decision-making group including the name of group and the decision problem through the web. The server checks the group's name assigned by the group leader. If the group name is valid, the server registers the group in the database and sends an approval to the client side. Other group members can then log in and register on the group through the web.

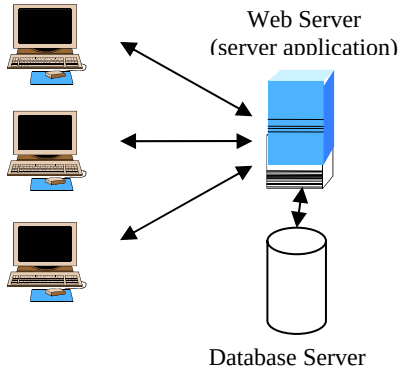


Fig. 2: Architecture of WFGDSS

The alternatives for the decision problem need to be stored into the database of WFGDSS before all members log in. After the group is set up the alternatives will be fetched from the database server and sent to client side by the server application. Based on these alternatives, each group member including the group leader proposes one or more assessment-criteria for selecting an alternative as the group satisfactory solution. All proposed assessment-criteria are then collected by the server application.

Referring to the assessment-criteria received from the server application, the group leader chooses top-T criteria as assessment-criteria for the decision problem in the group. As group members play different roles, the leader will assign weights, described by linguistic terms, to all group members. All data about top-T assessment-criteria and member's weights will be sent to the server, and then the database server for storage.

Based on the assessment-criteria and alternatives received, each group member is required to fill up a pairwise comparison matrix of the relative importance of these criteria and a belief level matrix to express the possibility of selecting a solution under some criteria. Once group members' two matrices are received, the server application first corrects the inconsistence of each pairwise comparison matrix of assessment-criteria based on linguistic inference rules, then calculate the belief level matrices, the belief vector, the normalized weights of group members, the weighted normalized fuzzy decision vector and the closeness coefficients of all alternatives consecutively. Finally, the web server constructs a final group decision page where the most satisfying group solution, which is corresponding to the maximum closeness coefficient, is displayed to all the group members.

B. An application of the WFGDSS

An executive group of a tourism company tries to determine which IT consulting firm to be hired in order to develop its e-tourism system. The main objectives to develop an e-tourism system are to present the company globally, build more interactive relationships with business partners and tourists, and reduce the costs of communication and market development. Four IT consulting firms have offered the e-tourism development services and each has submitted an

e-tourism system development proposal. Each firm and its proposal have advantages and disadvantages. The four firms' development proposals S_1, S_2, S_3 and S_4 are as alternatives for the tourism company. The executive group consists of three members P_1, P_2 and P_3, and P_1 is the leader. The three members have different opinions for selecting which firm to take the work and how to select one. The group must evaluate each firm's proposal by considering how to meet the company's objectives through the development of an e-tourism system.

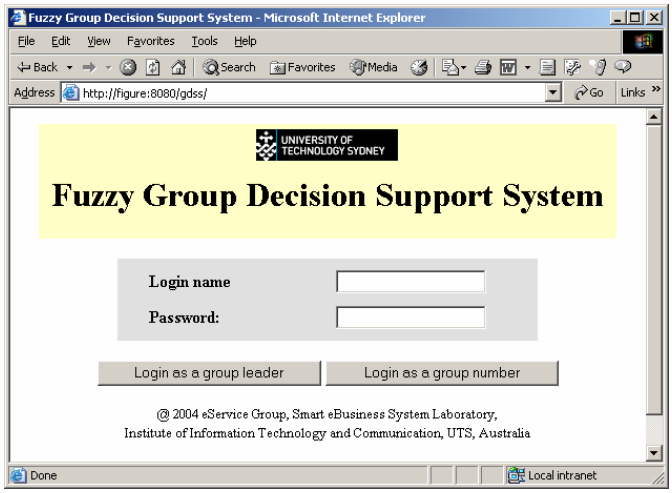


Fig. 3: Page for group member to log in

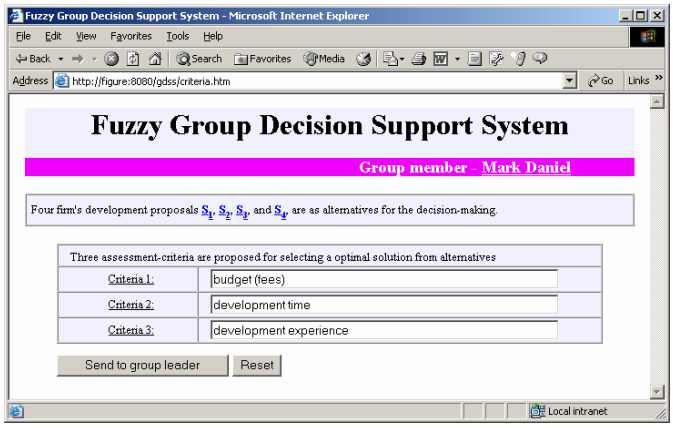


Fig. 4: Page for group member to propose assessment-criteria

Step 1: First of all, a group leader logs in to the system and defines a decision-making group as shown in Fig. 3. All other group members then join the group. Based on the four proposals, the three group members propose a number of assessment-criteria. For instance, a group member proposes budget (fees), development time and development experience as assessment-criteria for selecting a satisfactory firm from the four candidates which is shown as Fig. 4. The group leader collects all criteria as assessment-criteria and selects five: budget (fees), development time, development experience, quality of content support and quality of technique support as shown in Fig. 5.

Step 2: The group leader assigns weight "Strongly important person" to himself and "General decision maker" to other group members as shown in Fig. 5.

Step 3: Each member gives individual judgment for the five assessment-criteria by using AHP method. One group member's pairwise comparison matrix data is as shown in Fig. 6.

Step 4: Each group member gives a belief level of the possibility of selecting a solution under a criterion. One group member's belief level matrix data is shown in Fig. 6 as well.

By using linguistic inference rules, new comparison matrix and belief level matrix are shown as in Fig. 7.

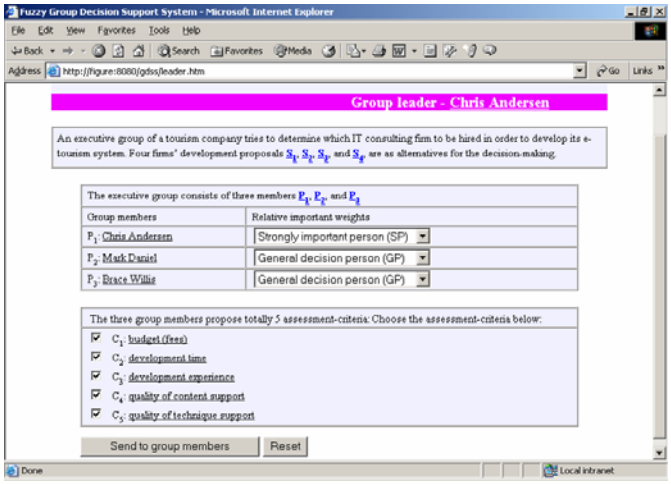


Fig. 5: Page for group leader to assign group members' weights and to choose the assessment-criteria

Step 5 –Step 9: After a series of calculation on belief vector, the weighted normalized fuzzy decision vector and the closeness coefficients of alternatives, Fig. 8 shows the closeness coefficient of all candidates and indicates the second one is the highest. That is, the second consulting firm is selected by the executive group.

The final group decision is the most acceptable by the group of individuals as a whole. A preliminary experiment has show that the model is appropriate for various multi-attribute decision problems, and can improve a group decision-making process and aid in functioning of a decision group.

V. Conclusion

This study first proposes a rational-political group decision-making model which carries out the advantages from both rational and political models, and therefore can handle inconsistent assessment, incomplete information and inaccurate opinions under a logical and sequential framework to get the best solution for a group decision. Based on the model, this study presents a linguistic term oriented fuzzy group decision-making method which allows group members to express their power, favor and judgment by linguistic terms. The method can use any type of fuzzy numbers to describe these linguistic terms. The method also uses inference rules to check preference consistence of each individual. The satisfactory group decision is derived as the most acceptable one for the decision group. It is very flexible and suitable for various group decision situations where alternatives are available. The method has been implemented by developing a

web-based group decision support system, called WFGDSS, where the web is as a development and delivery platform. Group members can use the WFGDSS asynchronously or synchronously, and don't need any training. In particular, the WFGDSS can be embedded into existing e-business or e-government systems through simple specification.

ACKNOWLEDGMENT

This research is supported by Australian Research Council (ARC) under discovery grant DP0211701.

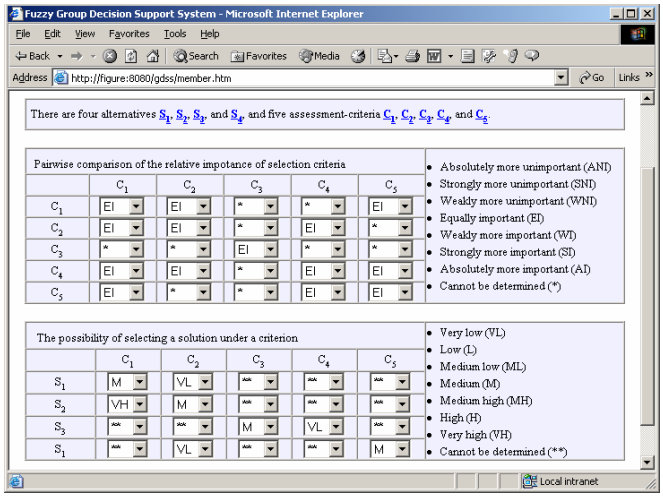


Fig. 6 Page for group member to input comparison of assessment-criteria and preferences

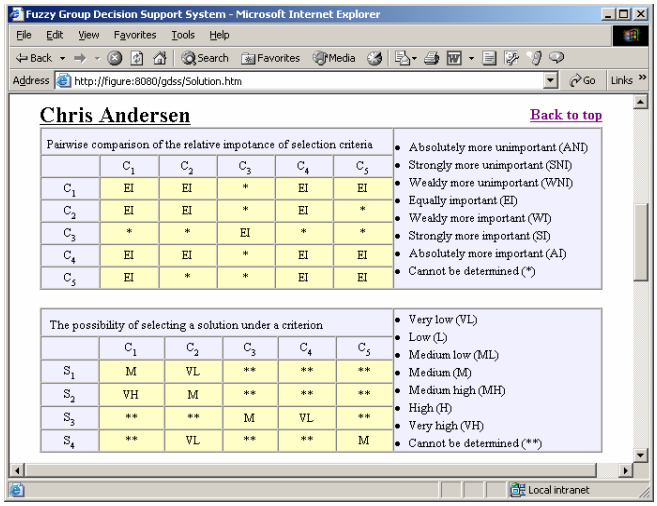


Fig. 7: Data for comparison of assessment-criteria and preferences after using linguistic inference

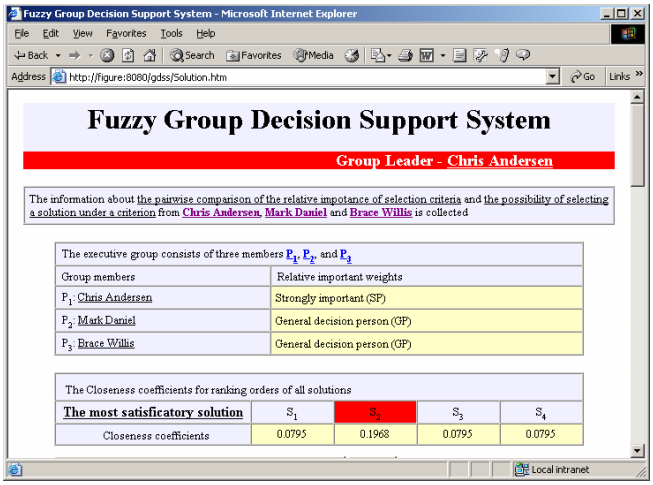


FIG. 8: PAGE FOR DISPLAYING CLOSENESS COEFFICIENTS AND THE GROUP SATISFACTORY SOLUTION

REFERENCES

[1] T. Bui, *Co-oP: A group decision support system for cooperative multiple criteria group decision-making*. Berlin: Springer-Verlag, 1989.

[2] G. Q. Zhang and J. Lu, "Chapter 3: Using general fuzzy number to handle uncertainty and imprecision in group decision-making", in *Intelligent Sensory Evaluation: Methodologies and Applications*, R. a. Zeng, Ed.: Springer, 2004, pp. 51-70.

[3] M. Marimin, I. Hatono, and H. Tamura, "Linguistic labels for expressing fuzzy preference relations in fuzzy group decision making", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 28, pp. 205-217, 1998.

[4] L. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning - part 1", *Information Sciences*, vol. 8, pp. 199-249, 1975.

[5] R. H. Sprague and H. J. Watson, *Decision Support Systems: Putting Theory into Practice, 2nd ed.* London: Prentice-Hall, 1989.

[6] M. Fedrizzi, J. Kacprzyk, and S. Zadrony, "An interactive multi-user decision support system for consensus reaching processes using fuzzy logic with linguistic quantifiers", *Decision Support Systems*, vol. 4, pp. 313-327, 1988.

[7] S. Sridhar, "Decision Support Using the Intranet", *Decision Support Systems*, vol. 23, pp. 19-28, 1998.

[8] K. J. Wang and C. F. Chien, "Designing an Internet-based group decision support system", *Robotics and Computer-Integrated Manufacturing*, vol. 19, pp. 65-77, 2003.

[9] D. J. Power and S. Kaparthi, "Building Web-based Decision Support Systems", *Studies in Informatics and Control*, vol. 11, pp. 291-302., 2002.

[10] U. Bose, A. M. Davey, and D. L. Olson, "Multi-attribute Utility Methods in Group Decision Making: Past Applications and Potential for Inclusion in GDSS", *Omega*, vol. 25, pp. 691-706, 1997.

[11] R. K. Lahti, "Group decision making within the organization: can models help?" CSWT papers, http://www.eorkteams.unt.edu/reports/lahti.htm, 1996.

[12] M. A. Lyles and H. Thomas, "Strategic problem formulation: biases and assumptions embedded in alternative decision-making models", *Journal of Management Studies*, vol. 25, pp. 131-145, 1988.

[13] J. Kacprzyk, M. Fedrizzi, and H. Nurmi, "Group decision making and consensus under fuzzy preference and fuzzy majority", *Fuzzy Sets and Systems*, vol. 49, pp. 21-31, 1992.

[14] I. Nishizaki and F. Seo, "Interactive support for fuzzy trade-off evaluation in group decision-making", *Fuzzy Sets and Systems*, vol. 68, pp. 309-325, 1994.

[15] G. Q. Zhang and J. Lu, "An integrated group decision making method with fuzzy preference for alternatives and individual

judgments for selection criteria", *Group Decision and Negotiation*, vol. 12, pp. 501-515, 2003.

[16] H. Hsu and C. Chen, "Aggregation of fuzzy opinions under group decision-making", *Fuzzy Sets and Systems*, vol. 55, pp. 241-253, 1996.

[17] N. Karacapilidis and C. Pappis, "Computer-supported collaborative argumentation and fuzzy similarity measures in multiple criteria decision-making", *Computers and Operations Research*, vol. 27, pp. 653-671, 2000.

[18] S. Chen, M., "Aggregation fuzzy opinions in the group decision-making environment", *Cybernetics and Systems*, vol. 29, pp. 363-376, 2000.

[19] C. T. Chen, "Extensions of the TOPSIS for group decision-making under fuzzy environment", *Fuzzy Sets and Systems*, vol. 114, pp. 1-9, 2000.

[20] H. Lee, "Group decision-making using fuzzy set theory for evaluating the rate of aggregative risk in software development", *Fuzzy Sets and Systems*, vol. 80, pp. 261-271, 1996.

[21] R. C. W. Kwok, J. Ma, and D. Zhou, "Improving group decision making: A fuzzy GSS approach", *IEEE Transactions on Systems, Man, and Cybernatics - Part C: Applications and Reviews*, vol. 32, pp. 54-63, 2002.

[22] J. P. Shim, M. Warkentin, J. F. Courtney, D. J. Power, R. Sharda, and C. Carlsson, "Past, Present, and Future of Decision Support Technology", *Decision Support Systems*, vol. 33, pp. 111-126, 2002.

[23] R. L. Lang and A. S. Whinston, "A Design of a DSS Intermediary for Electronic Markets", *Decision Support Systems*, vol. 25, pp. 181-197, 1999.

[24] H. Bhargava, R. Krishnan, and R. Muller, "Decision support on demanded: emerging electronic markets for decision technologies", *Decision Support Systems*, vol. 19, pp. 193-214, 1997.

[25] S. Ba, R. Kalakota, and A. B. Whinston, "Using Client--broker--server Architecture for Intranet Decision Support", *Decision Support Systems*, vol. 19, pp. 171-192, 1997.

[26] K. Poh, "Knowledge-based Guidance System for Multi-attribute Decision Making", *Artificial Intelligence in Engineering*, vol. 12, pp. 315-326, 1998.

[27] I. U. Sikder and A. Gangopadhyay, "Design and Implementation of a Web-based Collaborative Spatial Decision Support System: Organizational and Managerial Implications", *Information Resources Management Journal*, vol. 15, pp. 33-47, 2002.

RELATED CONFERENCES, CALL FOR PAPERS/PARTICIPANTS

TCII Sponsored Conferences

AWIC'05
The Third Atlantic Web Intelligence Conference
Lodz, Poland
June 6-9, 2005
http://wic.ics.p.lodz.pl/awic/

The 3rd Atlantic Web Intelligence Conference (Madrid - 2003, Cancun - 2004) brings together scientists, engineers, computer users, and students to exchange a nd share their experiences, new ideas, and research results about all aspects (theory, applications and tools) of intelligent methods applied to Web based systems, and to discuss the practical challenges encountered and the solutions adopted.

The conference will cover a broad set of intelligent methods, with particular emphasis on soft computing. Methods such as (but not restricted to):

Neural Networks, Fuzzy Logic, Multivalued Logic, Rough Sets, Ontologies, Evolutionary Programming, Intelligent CBR, Genetic Algorithms, Semantic Networks, Intelligent Agents, Reinforcement Learning, Knowledge Management, etc.

must be related to applications on the Web like:

Web Design, Information Retrieval, Electronic Commerce, Conversational Systems, Recommender Systems, Browsing and Exploration, Adaptive Web, User Profiling/Clustering, E-mail/SMS filtering, Negotiation Systems, Security, Privacy, and Trust, Web-log Mining, etc.

WI 2005
The 2005 IEEE/WIC/ACM International Conference on Web Intelligence
Compiègne, France
September 19-21, 2005
http://www.comp.hkbu.edu.hk/WI05/

Web Intelligence (WI) has been recognized as a new direction for scientific research and development to explore the fundamental roles as well as practical impacts of Artificial Intelligence (AI) (e.g., knowledge representation, planning, knowledge discovery and data mining, intelligent agents, and social network intelligence) and advanced Information Technology (IT) (e.g., wireless networks, ubiquitous devices, social networks, wisdom Web, and data/knowledge grids) on the next generation of Web-empowered products, systems, services, and activities. It is one of the most important as well as promising IT research fields in the era of Web and agent intelligence.

The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05) will be jointly held with The 2005 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'05). The IEEE/WIC/ACM 2005 joint conferences are sponsored and organized by IEEE Computer Society Technical Committee on Intelligent Informatics (TCII), Web Intelligence Consortium (WIC), and ACM-SIGART.

Following the great successes of WI'01 held in Maebashi City, Japan , WI'03 held in Halifax, Canada, and WI'04 held in Beijing, China. WI 2005 provides a leading international forum for researchers and practitioners (1) to present the state-of-the-art of WI technologies; (2) to examine performance characteristics of various approaches in Web-based intelligent information technology; and (3) to cross-fertilize ideas on the development of Web-based intelligent information systems among different domains. By idea-sharing and discussions on the underlying foundations and the enabling technologies of Web intelligence, WI 2005 will capture current important developments of new models, new methodologies and new tools for building a variety of embodiments of Web-based intelligent information systems.

IAT 2005
The 2005 IEEE/WIC/ACM International Conference on Intelligent Agent Technology
Compiègne, France
September 19-21, 2005
http://www.comp.hkbu.edu.hk/IAT05/

The 2005 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'05) will be jointly held with The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05). The IEEE/WIC/ACM 2005 joint conferences are sponsored and organized by IEEE Computer Society Technical Committee on Intelligent Informatics (TCII), Web Intelligence Consortium (WIC), and ACM-SIGART. The upcoming meeting in this conference series follows the great success of IAT-99 held in Hong Kong in 1999, IAT-01 held in Maebashi City, Japan in 2001, IAT-03 held in Halifax, Canada, and IAT-04 held in Beijing, China.

IAT 2005 provides a leading international forum to bring together researchers and practitioners from diverse fields, such as computer science, information technology, business, education, human factors, systems engineering, and robotics, to (1) examine the design principles and performance characteristics of various approaches in intelligent agent technology, and (2) increase the cross-fertilization of ideas on the development of autonomous agents and multi-agent systems among different domains. By encouraging idea-sharing and discussions on the underlying logical, cognitive, physical, and sociological foundations as well as the enabling technologies of intelligent agents, IAT 2005 will foster the development of novel paradigms and advanced solutions in agent-based computing.

ICDM'05
The Fifth IEEE International Conference on Data Mining
New Orleans, Louisiana, USA
November 26-30, 2005
http://www.cacs.louisiana.edu/∼icdm05/
Submission Deadline: June 15, 2005

The 2005 IEEE International Conference on Data Mining (IEEE ICDM '05) provides a premier forum for the dissemination of innovative, practical development experiences as well as original research results in data mining, spanning applications, algorithms, software and systems. The conference draws researchers and application developers from a wide range of data mining related areas such as statistics, machine learning, pattern recognition, databases and data warehousing, data visualization, knowledge-based systems and high performance computing. By promoting high quality and novel research findings, and innovative solutions to challenging data mining problems, the conference seeks to continuously advance the state of the art in data mining. As an important part of the conference, the workshops program will focus on new research challenges and initiatives, and the tutorials program will cover emerging data mining technologies and the latest developments in data mining. technologies and the state-of-the-art of data mining developments.

Topics related to the design, analysis and implementation of data mining theory, systems and applications are of interest. See the conference Web site for more information.

Related Conferences

AAMAS'05
The Fourth International Joint Conference on Autonomous Agents and Multi-Agent Systems
Utrecht, The Netherlands
July 25-29, 2005
http://www.aamas2005.nl/

AAMAS-05 encourages the submission of theoretical, experimental, methodological, and applications papers. Theory papers should make clear the significance and relevance of their results to the AAMAS community. Similarly, applied papers should make clear both their scientific and technical contributions, and are expected to demonstrate a thorough evaluation of their strengths and weaknesses in practice. Papers that address isolated agent capabilities (for example, planning or learning) are discouraged unless they are placed in the overall context of autonomous agent architectures or multi-agent system organization and performance. A thorough evaluation is considered an essential component of any submission. Authors are also requested to make clear the implications of any theoretical and empirical results, as well as how their work relates to the state of the art in autonomous agents and multi-agent systems research as evidenced in, for example, previous AAMAS conferences. All submissions will be rigorously peer reviewed and evaluated on the basis of the quality of their technical contribution, originality, soundness, significance, presentation, understanding of the state of the art, and overall quality.

In addition to conventional conference papers, AAMAS-05 also welcomes the submission of papers that focus on implemented systems, software, or robot prototypes. These papers require a demonstration of the prototype at the conference and should include a detailed project or system description specifying the hardware and software features and requirements.

IJCAI'05
The Nineteenth International Joint Conferenceon on Artificial Intelligence
Edinburgh, Scotland
July 30 - August 5, 2005
http://ijcai05.csd.abdn.ac.uk/

The IJCAI-05 Program Committee invites submissions of full technical papers for IJCAI-05, to be held in Edinburgh, Scotland, 30 July - 5 August, 2005. Submissions are invited on substantial, original, and previously unpublished research on all aspects of artificial intelligence.

ISWC2004
The Fourth International Semantic Web Conference
Galway, Ireland
6-10 November, 2005
http://iswc2005.semanticweb.org/

ISWC is a major international forum at which research on all aspects of the Semantic Web is presented. ISWC2005 follows the 1st International Semantic Web Conference (ISWC2002 which was held in Sardinia, Italy, 9-12 June 2002), 2nd International Semantic Web Conference (ISWC2003 which was held in Florida, USA, 20 - 23 October 2003) and 3rd International Semantic Web Conference (ISWC2004 which was held in Hiroshima, Japan, 7-11 November 2004).

ICTAI2005
The Seventeenth IEEE International Conference on Tools with Artificial Intelligence
Hong Kong, China
14-16 November, 2005
http://ictai05.ust.hk/

The annual IEEE International Conference on Tools with Artificial Intelligence (ICTAI) provides a major international forum where the creation and exchange of ideas relating to artificial intelligence are fostered among academia, industry, and government agencies. The conference facilitates the cross-fertilization of these ideas and promotes their transfer into practical tools, for developing intelligent systems and pursuing artificial intelligence applications. The ICTAI encompasses all the technical aspects of specifying, developing, and evaluating the theoretical underpinnings and applied mechanisms of AI tools. A selection of the best papers in the conference will be published in a special issue of the International Journal on Artificial Intelligence Tools (IJAIT), special issues of other journals, or edited books.

IEEE Computer Society
1730 Massachusetts Ave, NW
Washington, D.C. 20036-1903

Non-profit Org.
U.S. Postage
PAID
Silver Spring, MD
Permit 1398