On-line Hot Topic Recommendation Using Tolerance Rough Set Based Topic Clustering

Yonghui Wu, Yuxin Ding, Xiaolong Wang, Jun Xu
Intelligence Computing Research Center
Harbin Institute of Technology, Shenzhen Graduate School, Shenzhen, People's Republic of China
{yhwu,wangxl}@insun.hit.edu.cn, yxding@hitsz.edu.cn, hit.xujun@gmail.com

***Abstract*—In this paper we present our research of online hot topic detection and label extraction method for our hot topic recommendation system. Using a new topical feature selection method, the feature space is compressed suitable for an online system. The tolerance rough set model is used to enriching the small set of topical feature words to a topical approximation space. According to the distance defined on the topical approximation space, the web pages are clustered into groups which will be merged with document overlap. The topic labels are extracted based on the approximation topical space enriched with the useful but high frequency topical words dropped by the clustering process. The experiments show that our method could generate more information abundant classes and more topical class labels, alleviate the topical drift caused by the non-topical and noise words.**

***Index Terms*—topic detection, tolerance rough set model, association rule, clustering, recommendation system**

I. INTRODUCTION

Internet users get information by inputting key words into search engines. However, lots of people are interested in the important news or topics in a specific domain which can not described by "key words". These users don't know the key words and don't need to input any thing. For example, the stock owners need a close following about the financial news to find the most potential stocks. But they don't know the "key words" before the related news come out of websites. News recommendation system can solve this problem.

In this research, we present a topic detection and news recommendation method, which could recommend the most important news and topics about a certain domain without the user logs. The user log-based news recommendation system needs both the user's data and the news data. This kind of recommendation is a filtering problem: using the user's data as a filter, which is widely used in e-commerce website where the past shopping records are used as a filter for new product recommendation. Our recommendation method could find the topics in web page collections, take out the topic labels, and recommend the most important topics to the user. This kind of recommendation is a detection problem. There are many researches about the topic detection and tracking (TDT). However, most of them are based on the offline corpus. The traditional TF-IDF based feature selection is not suitable in a topic detection process for the topic drift caused by the non-topical words and noise words. This topical drift can be alleviated by a tolerance rough set based topical approximation feature space. Most of the online topic research are focused on the new topic detect task. The topic label extraction is also an important part of news recommendation system. Association rules extraction is widely used in e-commerce recommendation system for extraction of recommendation rules from shopping-record, which is compact and short. The web page contains more noise words compared to the shopping record. When used in topic label extraction from web pages, the association rule methods also suffers the topic drifting problem caused by the non-topical words. When the number of pages grows up, the time complexity of association rule method increased sharply.

In this research, we present a tolerance rough set based topical clustering and class label extraction method based on a tolerance approximation topical feature space, which could reduce the non-topical noise words and generate more topical cluster. By controlling the threshold parameter the tolerance approximation topical feature space is more compact and topical than the TF-IDF feature space. The experiments show our method is suitable for an on-line recommendation system.

The remaining of the paper is organized as follows. In section 2, the related research works are given. Section 3 presents the discussion of web page crawling and testing corpus. Section 4 discusses the topic detect using tolerance rough set based clustering, the topic label extraction using association rules. The experiment results and future works are shown in section 5 and section 6, respectively.

II. PREVIOUS WORKS

A. Topic Detection and Website News Recommendation

The traditional TDT research defined a topic as "a seminal vent or activity, along with all directly related events and activities" [1]. In our news recommendation system, a topic is defined as a group of news pages reporting the same or related events. There are many TDT research based on the TDT corpus, which contains document labeled into different topics. Most of the TDT research using the information filtering methods. James Allan Ron papka [2] use the surprising features tracking new event from the Internet. Bingjun Sun and Prasenjit

doi:10.4304/jcp.5.4.549-556

Mitra [3] use mutual information (MI) and weighted mutual information (WMI) tracking the topics in multi-document environment. Abhinandan Das and Mayur Datar [4] use a collaborative Filtering method in the Google's personalized news recommendation system. Kuan-Yu Chen and Luesak Luesukprasert [5] make research of the hot topic extraction using sentence modeling and time-line analysis. Clustering methods [6] is a useful tool in the topic detection process.

B. Rough Set Theory

Rough set theory is used for approximation of a concept, which is an extension of the set theory. This theory is first introduced by Pawlk [7], and then widely used in mining the association between attributes. Rough set Theory is widely used in many researches such as: latent semantic analysis, expert system with incomplete information, Association rules extraction ... and so on. Several models based on rough set are proposed in the information retrieval.

Funakoshi and Ho [8] proposed Equivalence rough set model, which using the equivalence classes. This model can be used to calculate the latent semantic between terms, but it is hard to calculate the equivalence classes due to the strict transitive property restrict.

Skowron and Stepaniuk [9] proposed the tolerance rough set model (TRSM) in 1994. Instead of using the strict equivalent relation TRSM relax the transitive restriction which makes the TRSM feasible to use in Information retrieval.

C. Association Rules

The association rules extraction is widely used in the research of the transaction records containing compact data. J.Han and Y.Fu [10], R.Agrawal [11] deal with the large database association rule extraction problem. S. Wesly Changchien and Tzu-Chuen Lu [12] using a neural network based clustering and association rules extraction method in on-line products recommendation system. M.Delgado [13] present a fuzzy association rules extraction method in text mining based on fuzzy transaction.

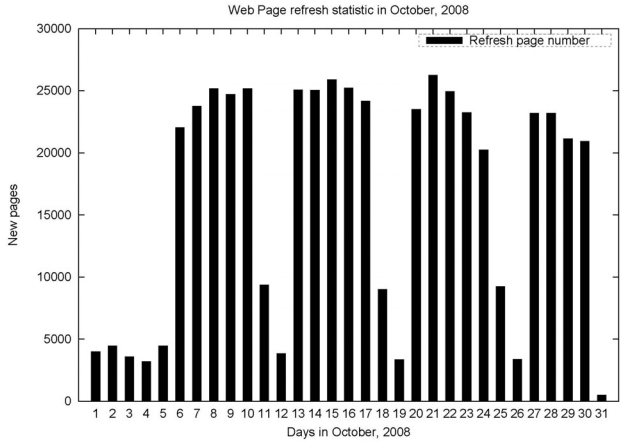
III. WEB PAGE COLLECTION

The web pages used in this research is collected by the crawler of our search engine Inar. In order to get the most useful web pages from a specific domain website a new edition is developed from the Inar crawler: Pooler hot news crawler. The detailed information about the crawler can be seen from [14] [15].

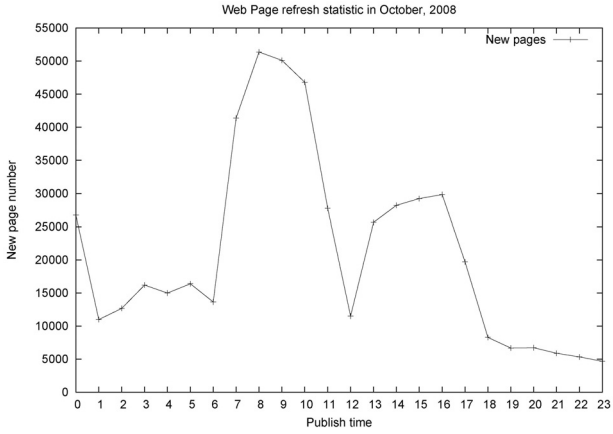
Pooler uses an improved aggregate ranking algorithm [16] to ranking the websites in a specific domain, find the most useful set as well as the index pages, refresh the page set incrementally.

In this research, we use the financial websites as an example. As an online topic recommendation system, we are focused on the new web pages refreshed everyday.

The Pooler collects about 10,650,000 pages from about 2600 financial website. The top 70 site are chosen for this research by the improved aggregate ranking method. The publish time of web pages are adjusted using the time label extracted from the page which could prevent the lost of "last-modify" part in many website servers. In order to get the number of pages processed by the recommendation system, we sampled the web pages with a publish time in October, 2008.



(a) New page refreshed daily



(b) New page refreshed hourly

Figure 1. Web page refreshing in October, 2008

TABLE I.
NEW PAGES REFRESHED IN OCTOBER, 2008

Number of Refreshing Pages	Days	Number of Websites
510806	30	70

The number of pages refreshed in October 2008 can be seen from Table I. Fig. 1(a) shows the new page refreshed by days in October, 2008. Fig. 1(b) shows the refreshing new pages by hours regardless of the date.

IV. TOPIC DETECTION IN APPROXIMATION SPACE

The clustering method always suffers the noise words. In topical clustering of web pages, we call the words not related with its topic as the non-topical words. Non-topical words could drift the topical clustering process. In order to get more topical clustering results, several methods are proposed to reduce the non-topical words as

well as the feature space. The reduction of the feature space may cause the lost of co-occurrence relation of words, which is an important feature in clustering.

In this research, we use a tolerance rough set based clustering method to get more topical class. After feature selection the non-topical words are deserted. This could generate a compact feature space suitable for an online recommendation system. Then this feature space is extended into a topical approximation space using the TRSM. By changing the TRSM parameters, the size of the topical approximation space can be controlled which could balance the clustering result and the running time. Then for each class the topic label is extracted using association rule on the tolerance approximation feature space.

A. Topical Feature Selection

In order to get a topical clustering result, the topical words sufficiently describe the topic must be extracted. Several term weighting schemes are used in selecting the key words from the web pages. Term frequency-inverse document frequency (TF-IDF) is widely used in text mining. The TF-IDF method can drop the high frequency words which are disaster for the clustering. However, the hot topics occurs in many web page will be dropped by this term weighting method. So we using different feature selection in the topic label extraction, which gives the topical words which occur frequently on many web sites a high score.

Another disadvantage of the TF-IDF as well as other frequency based feature selection is all the worlds are treated equally, which is not suitable for a topical clustering. The topical words are more important than other context in a topical cluster. If the feature space contains more topical words, the cluster result will be more topical.

Usually, the web page contains the “title” filed, which in most time is the key words of the content and should be given a high score than the content. The words close to or co-occur with the title words can be used to enrich the topical words.

At first, the web pages collected by crawler Pooler are purified by a cleaning module to drop the scripts and noise texts such as the advertisements. Some of the non-informative parts are dropped by this cleaning module, such as the numbers, the hyperlinks in text area, the error message generated by the script, the foot note of web pages. These no-topical words usually cause the topical drift in clustering process. For example, the web pages are usually generated from some predefined models. The pages collected from the same web site usually contain the same navigation text, copy right information, as well as the foot note information. The page co-occurrence of these texts is even higher than the topical words, which lead to a low NTD distance. This non-topical distance could be alleviated by web sited distance (NSD). However, these non-topical distances are low enough to drift the clustering process in many times.

Then the texts are processed by the Chinese word segmentation and Named Entity Recognition system ELUS [17]. The ELUS system uses a trigram model with smoothing algorithm for Chinese word segmentation and a Maximum entropy model for Named Entity Recognition. In the third SIGHAN-2006 bakeoff the ELUS system achieves F-measure 96.8%, which is the best in MSRA open test. The part of speech features can be extracted effectively using ELUS system.

The title words and the noun word of web pages are extracted as the initial topical words set.

B. Topical Words Enriching

There are few topical words derived from the title. More topical words can be found from the co-occur set of these topical words, which is the latent semantic features. Latent Semantic Analysis (LSA) method can be used to find the topical words relation. However this kind of method usually cost too much complex computing for an online system.

TRSM is another useful tools to mining the latent semantic feasible for on-line clustering. The original topical words set can be extend to a upper approximation topical words set in which the two web page are related to a topic even if they are not related in the co-occur feather space.

Tolerance Rough Set Model: Let U be the universe, $R \subseteq U \times U$ be an equivalence relation defined on U. A = (U, R) is an approximation space. In the original rough set theory the lower and upper approximation of set $X \subseteq U$ with respect to relation R is defined as:

$$\underline{R}[X] = \{x \in U : [x]_R \subseteq X\}. \quad (1)$$

$$\overline{R}[X] = \{x \in U : [x]_R \bigcap X \neq \varnothing\}. \quad (2)$$

Where $[X]_R = \{y \in U \mid xRy\}$ is the equivalence class of objects indiscernible with x regarding R. Then the rough boundary of X can be defined as:

$$BN_R(X) = \overline{R}(X) - \underline{R}(X). \quad (3)$$

A topic is always overlapped with other related topics, so we have to deal with the overlapping classes derived from the clustering process. The approximation of rough set theory is useful to deal with this kind of overlapping concept. The original small size of topical words set can be enriching using this approximation. However, the relation R in original rough set theory, which is reflective, symmetric and transitive are too strict for natural language processing (NLP). As a matter of fact, these transitive in topical clustering always lead to a topic drift.

Skowron and Stepaniuk proposed a tolerance rough set model (TRSM) by relaxing the transitive restrict. The restrict ion of the relation R in TRSM is reflexive and symmetric. The approximation space derived using TRSM is called tolerance space.

TRSM can be defined as a quadruple:

$$\Box = (U, I, v, P). \quad (4)$$

Which U is a non-empty set of objects, $I : U \to P(U)$, $v : P(U) \times P(U) \to [0,1]$, $P : I(U) \to \{0,1\}$ are uncertainty

vague inclusion and structure function where $P(I(x))=1$ for any $x \in U$. The lower and upper approximation in R for any $X \subset U$ can be defined as:

$$\underline{R}[X]=\{x \in U : P(I(x))=1 \wedge v(I(x),X)=1\}. \quad (5)$$

$$\overline{R}[X]=\{x \in U : P(I(x))=1 \wedge v(I(x),X)>0\}. \quad (6)$$

The relation R in TRSM for text information retrieval is usually defined as the co-occurrence in document set D. In our topical feature approximation the R is defined as the topical co-occurrence in all document set D. We tried many ways to define the topical co-occurrence. In this paper we define the topical co-occurrence as the words co-occurred in the topical area. The web page title is an important topical area to generate the original topical word set (TWS). The topical area is defined as the text area around the topical word in TWS with a distance of l . If the page title contains only few topical words not enough to describe the document the part of speech (POS) feature is used to enrich the feature set.

Let $D=\{d_1,d_2,\cdots,d_n\}$ be a set of documents, and $T=\{t_1,t_2\cdots,t_m\}$ be a word set co-occurred in the topical area of D. The topical approximation space $R=\{T,I_\theta,v,P\}$ can be defined over T using uncertainty function I_θ :

$$I_\theta(t_i)=\{t_j \mid f_{TO}(t_i,t_j) \geq \theta\} \bigcup \{t_i\}. \quad (7)$$

Where θ is a user defined threshold, $f_{TO}(t_i,t_j)$ is the topical co-occurrence function of words t_i and t_j in distance l. The lower and upper topical approximation can be defined as:

$$\underline{R}[X]=\{t_i \in T : v(I_\theta(t_i),X)=1\}. \quad (8)$$

$$\overline{R}[X]=\{t_i \in T : v(I_\theta(t_i),X)>0\}. \quad (9)$$

The original topical word set can be derived from the web page title as well as the POS feature words. The upper topical approximation $\overline{R}$ is a set of words which related to the topic in the topical approximation feature space.

C. Topical Clustering

Semantic Distance Measurement of Topical words: The measurement of word distance is an important factor of the clustering method. There are many ways to compute the semantic distance between two topical words.

The manually labeled dictionary such as WordNet and synonym word tables can be used in distance measuring. This kind of manually label based tools suffers many disadvantages in topical clustering. First, the inconsistent between the manually labeled relations and the new arriving web pages usually lead to a topic drift in the topic clustering process. The Internet has no dearth of content. Every time there comes thousands of new web pages contains topical words inconsistent with the manually labeled relations. Second, the commonly used WordNet like lexicons has rare sense when used in specific domains. Maintaining of labeled lexicons of a specific domain is not easy.

Some statistical based method is proposed based on the assumption that the two words tend to be related if they have a high co-occurrence. This method makes good use of the specific word co-occurrence features in web pages. The Normalized Google Distance (NGD) is used in many related papers. The NGD method uses the number of pages containing the words returned by Google as the measurement of the distance between two words. The normalized Google distance between words w_1 , w_2 can be defined as :

$$NGD(w_1,w_2)=\frac{\max\{\log f(w_1),\log f(w_2)\}-\log f(w_1,w_2)}{\log N-\min\{\log f(w_1),\log f(w_2)\}}. \quad (10)$$

Where $f(w_i)$ are documents contains word w_i , N is the total number of the document set D, $f(w_1,w_2)$ is the number of web pages containing word w_1 and w_2 . Words with high topic relation are tending to have low Google distance.

The NGD distance can be derived using the Google engine by counting the page numbers returned from Google. However the topical approximation set is large and changing every time the new page comes which make it not feasible in practice. In this research, we use a page set from a specific domain to compute the distance like the NGD. This web page collection comes from the specific domain contains specific features which can not be extracted from a common web page collection.

The NGD like distance is generated by the page co-occurrence of words. The web pages have more abundant information than text files, such as the source website names. At the beginning of this research we use the NGD like distance derived from our web page collections. We find that the topical words co-occurred in many websites is more important than the words co-occurred in many web pages belongs to the same website. In another word, the topical words occur in many websites (high website co-occurrence words), is more important than the words with high page co-occurrence.

The source website name is a useful feature which could give us more information about the distance between two topical words. The website frequency of topical words can be derived from the URL.

The distance used in this research is the Normalized Topical Distance (NTD), which composed of the Normalized Page Distance (NPD) and the Normalized Site Distance (NSD).

$$NPD(w_1,w_2)=\frac{\max\{\log f_P(w_1),\log f_P(w_2)\}-\log f_P(w_1,w_2)}{\log N_P-\min\{\log f_P(w_1),\log f_P(w_2)\}}. \quad (11)$$

$$NSD(w_1,w_2)=\frac{\max\{\log f_S(w_1),\log f_S(w_2)\}-\log f_S(w_1,w_2)}{\log N_S-\min\{\log f_S(w_1),\log f_S(w_2)\}}. \quad (12)$$

Where $f_P(w_i)$ and $f_S(w_i)$ are respectively the number of pages contain word w_i and the number of websites contain word w_i. N_P is the total number of web pages. N_S is the total number of websites.

Then we can define our NGD like distance NTD as follows:

$$NTD = \alpha_P \times NPD + \alpha_S \times NSD . \quad (13)$$

The parameter α_P and α_S are used to control the weight of NPD and NSD. By this definition, if two topical words group co-occurred in the same number of pages but deferent number of websites, the distance between two words co-occurred in more websites are tends to be closer.

The NSD between two topical words can be computed along with the NPD computing process. The computing complexity of NTD is the same with traditional NGD.

The NGD like method suffers a bias when two words co-occurred only with each other. Since low document frequency words are useless in clustering process, a threshold is used to drop these words. The high document frequency words are also dropped by an upper threshold.

Topical Clustering Process: After the topical words selection all the web pages are presented by the topical words set. These topical words are clustered into groups that each group contains the topical words highly co-related in the topical approximation space using complete-link clustering. The complete-link clustering tend to produce tightly bounded compact clusters than the single-link clustering, which is helpful to generate as much topic as the topical approximation space contains.

Cluster Merging: The words from the same topic may be clustered into different groups due to the tightly bounded clusters generated by the complete-link clustering. There are too many groups to be integrated into our hot news recommendation system. As we consider the NSD in our distance measurement, the topical words with both high page co-occurrence and website co-occurrence tends to have low NTD, which makes these worlds more likely to be clustered into a small class with only several topical words. These groups should be merged using other topical relation measure method. The document overlap of each class can be used to deal with these small classes.

All the clusters generated by the first clustering process are merged by the document overlap between them. The overlap between two class C_1 and C_2 is defined as:

$$OL(C_1, C_2) = \frac{|C_1 \bigcap C_2|}{MIN\{|C_1|, |C_2|\}} . \quad (14)$$

Where $|C_i|$ is the document number in group C_i. In the experiments the threshold of document overlap is set to 0.85.

D. Topical Label Generation

The topical label generation requires more topical related label. In feature selection process the topical words with high document frequency is dropped, which may be the topical words. The label generated by traditional cluster label extraction method can not well describe the topical meanings.

Using association rules method we can generate topical labels suitable for on-line news recommendation system.

The association rule extraction is proved to be effective in recommendation system based on DB or other short records. These kinds of records are compact and commonly contain few words, such as the shopping record. The web pages, however, contains a large set of words which cost more computing time. The computing cost could be alleviated by using the association rules extraction algorithm in each class generated by the topical clustering.

The output of the topic label generation is a group of topical words which is the hot topical words. The first step is select candidate hot topical words from the web page of each class.

The topical words sets are selected in the topical clustering process. But some high frequency topical words are dropped by the feature selection. In topic label generation we select all the high frequency noun words into the candidate topical words set.

The hotness $H(W_i)$ of each topical world W_i is defined as:

$$H(W_i) = \lambda_1 \times \frac{Site(W_i)}{|S|} + \lambda_2 \times \frac{Page(W_i)}{|P|} . \quad (15)$$

Where $Site(W_i)$ is the number of website which contains a web page related to topical word W_i, S is the total number of website select by the ranking component, $Page(W_i)$ is the number of web pages containing topical words W_i, P is the total number of pages.

Input: Candidate topic word set $W_1 \ldots W_n$,
Support threshold θ_s, Cover threshold θ_c
Output: Topic Label L
$L \Leftarrow \varnothing$;
for *each* W_i **do**
  $L = L \cup \{W_i\}$
end
while 1 **do**
  $T \Leftarrow \varnothing$;
  for $l_i, l_j \in L; i \neq j;$ **do**
    if $SUP(l_i, l_j) \geq \theta_s$ **then**
      $T = T \cup \{l_i \cup l_j\}$;
    end
  end
  for $t_i, t_j \in T; i \neq j;$ **do**
    if $Cover(t_i, t_j) \geq \theta_c$ **then**
      $T = (T - t_i - t_j) \cup (t_i \cup t_j)$;
    end
  end
  if $T == \varnothing$ **then**
    break;
  end
  else
    $L \Leftarrow T$;
  end
end
return L ;

Algorithm 1: Topic Label Generation Process

As the traditional association rules extraction compute the support value between each record, we compute the support between each candidate hot topical words. The support of two topical word set WS_1 and WS_2 can be calculated as follow:

$$SUP(WS_1, WS_2) = \lambda_1 \times \frac{Site(WS_1 \bigcap WS_2)}{\max\{|Site(WS_1)|, |Site(WS_2)|\}} + \lambda_2 \times \frac{Page(WS_1 \bigcap WS_2)}{\max\{|Page(WS_1)|, Page(WS_2)\}} \quad (16)$$

Where the support of two topical word set composed of two parts: the website frequency support as well as the document frequency support. Parameter λ_1 and λ_2 can be used to control the hot weight caused by website and page. In this research λ_1 is set to 0.75 and λ_2 is set to 0.25. Taking the top N topical words as the candidate hot word, the topic label generation can be described as Algorithm 1. In the topic label generation, each of the hot topical word is a member of the label set L. Then the support between each $l_i, l_j \in L$ is computed. If the support value reaches the threshold the union set of l_i, l_j becomes a new member of the label set.

After each supporting compute loop, the new members in T will be merged when the overlap of the two members set reach the cover threshold θ_c.

The final label set for this class is the one contains the most of the topical words.

V. EXPERIMENTS AND RESULTS

This research is designed to improve the financial news recommendation system, which is an important part of the haitianyuan knowledge service platform. The topics are ranked by the number of topical words as well as the number of pages. Then the system returns the most important page titles derived from the top topics without the label generation process. With the topical clustering method in this research our recommendation system will be able to recommend hot topics of current financial website which is more useful than the hot titles.

Usually, the user's experience is the most useful tools in the evaluation of a clustering algorithm, which makes the comparing of different clustering algorithms a non-trivial task.

A. En vironment and Testing Corpus

The recommendation system is implemented using c++ and PHP on Linux system. The crawler is running on a Linux machine shows in table II. The web page corpus is randomly select from the news pages refreshed in October, 2008. The non-informative pages are deserted.

B. Minimum Word Frequency

Minimum word frequency threshold θ_{wf} is very important for the topical clustering. The low threshold will takes more words come into the topical word set, which makes the clustering generate more basic class and has a high Coverage.

However, the time cost is growing with the class number. The noise word in the TF-IDF based feature selection will generate many useless class and the non-topical words will drift the topical clustering process.

Fig. 2 shows the relation between the minimum term frequency threshold and the base class number in our topical approximation space compared with TF-IDF feature space in a corpus containing 2500 web pages. The threshold distance used in clustering is set to 0.3. Both methods will get a small number of classes using a high threshold and a large number of classes using a low threshold. The relation between threshold and class coverage can be seen in Fig. 3. Both methods get a high coverage in low threshold and a low coverage in high threshold.

As can be seen from Fig. 3, the POS feature and tolerance rough set based approximation feature space could drop the non-topical relations between non-topical words; reduce the number of class and running time, which is suitable for online using. The class coverage of topical approximation feature space is decrease for the non-topical words are dropped in feature selection.

C. Time Cost

In order to test the time complexity of this algorithm, we sampled sets containing different pages from 500 to 5000. Fig. 4 shows the time cost line for different numbers of web page sets which is near to linear. As we can see from Fig. 1(a), there are 5000 to 25000 new web pages refreshed every day. The page number can be controlled by discarding the redundant and topical less pages to get a reasonable time cost for a 24-hour on-line hot topic recommendation system.

VI. CONCLUSION AND FUTURE WORK

In this paper, a tolerance rough set based clustering method is presented to detect the topics in financial web site news pages. In the feature selection we use the part of speech feature and the title words to form initial topical word set. Then, this topical word set are extended by the TRSM using the co-occurrence of topical words in distance l. Based on the NGD method, the web site distance (NSD) is introduced into the new NGD like distance measure method: NTD. Using NTD distance the news pages are clustered into more topical class than the traditional feature space. The class label is generated by the association rule method.

The experiments show that the tolerance rough set based topical approximation space can exploit the latent semantic under the co-occurrence of topical words. This method alleviates the non-topical word caused topical drifting, reduces the non-topical class numbers as well as the running time. The topical clustering and label generation method could generate more topical class than the traditional TF-IDF features.

For the application of this method, there are some future tasks as how to detect the web pages in which the title inconsistent with the content, finding the most important topics from the clustering result. The title words of web pages are treated important in our method,

which is a good topical feature for the web pages with page titles consistent with the page contents. For the web pages with title inconsistent with the contents, we must find method to find the real topical words set.

Part of this research is integrated into our hot financial news detection system (http://news.haitianyuan.com.cn), which is an important part of knowledge service plat form haitianyuan.

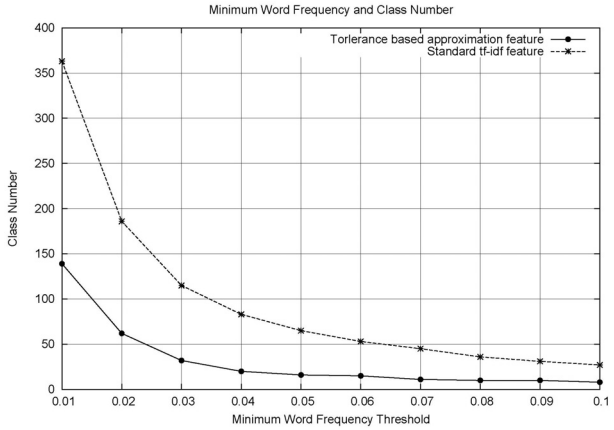


Figure 2. Minimum word frequency threshold and class numbers

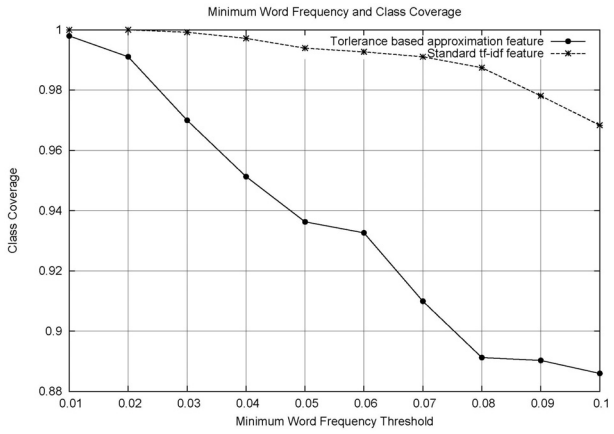


Figure 3. Minimum world frequency threshold and class coverage

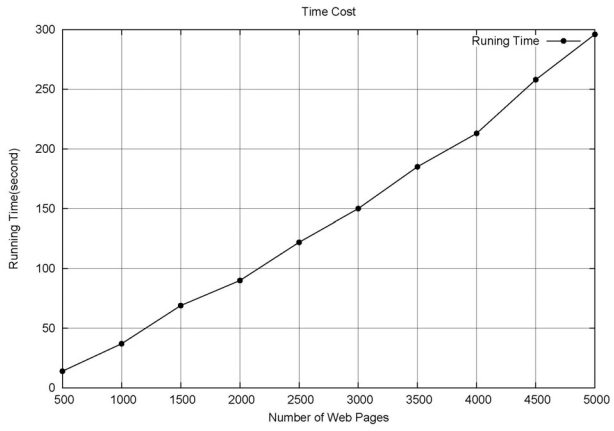


Figure 4. Running time for different web page numbers

TABLE II.
SYSTEM RUNING ENVIRONMENT

CPU	Intel Pentum 3.0
Memory	2G
Disk	320G

ACKNOWLEDGMENT

The authors wish to thank Dr. Xianjun Meng for his helpful suggestions.

This investigation was supported in part by the National Natural Science Foundation of China(No. 60703015) and the national 863 Program of China(No. 2006AA01Z197, No. 2007AA01Z194).

REFERENCES

[1] TDT2004, "The 2004 topic detection and tracking task definition and evaluation plan," 2004.
[Online]. Available: http://www.nist.gov/speech/tests/tdt/

[2] J. Allan, R. Papka, and V. Lavrenko, "On-line new event detection and tracking," SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 37–45, November 1998.

[3] B. Sun, P. Mitra, H. Zha, C. Giles, and J. Yen, "Topic segmentation with shared topic detection and alignment of multiple documents," SIGIR 07, pp. 23–27, July 2007.

[4] A. S. Das, M. Datar, A. Garg, and S. Rajaram, "Google news personalization: scalable online collaborative filtering," in WWW '07: Proceedings of the 16th international conference on World Wide Web. New York, NY, USA: ACM, 2007, pp. 271–280.

[5] K.-Y. Chen, L. Luesukprasert, and S. cho T.Chou, "Hot topic extraction based on timeline analysis and multidimensional sentence modeling," IEEE Transactions on Knowledge and Data Engineering, pp. 1016–1025, August 2007.

[6] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," ACM Computer Survey, vol. 31, no. 3, pp. 264–323, 1999.

[7] Z. Pawlak, "Rough sets," International Journal of Computer and Information Sciences. 11(5), pp. 341–356, 1982.

[8] K. Funakoshi and T. B. Ho, A Rough Set Approach to Information Retrieval. In: L. Polkowski and A. Skowron, (eds.) Rough Sets in Knowledge Discovery. Erewhon, NC: Physica-Verlag, 1998.

[9] A. Skowron and J. Stepaniuk, "Tolerance approximation space," 3rd International Workshop on Rough Sets and Soft computing, pp. 156– 163, November 1994.

[10] J. Han and Y. Fu, "Discovery of multiple-level association rules from large databases," in Proceedings of the 21st VLDB Conference, 1995, pp. 1–12.

[11] R. Agrawal, T. Imieli´nski, and A. Swami, "Mining association rules between sets of items in large databases," in SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data. New York, NY, USA: ACM, 1993, pp. 207–216.

[12] S. W. Changchien and T. Lu, "Mining association rules procedure to support on-line recommendation by customers and products fragmentation," Expert System with Applications, vol. 20, pp. 325–335, 2001.

[13] M.Degado, M. Martin-Bautista, D. Sanchez, and J.M.Serrano, "Association rule extraction for text mining," in Lecture Notes in Computer Science, 2002, pp. 154–162.

[14] Y. Ding, X. Wang, L. Lin, and Y. Wu, "The design and implementation of the crawler-inar," ICMLC, pp. 13–16, August 2006.

[15] Y. Wu, X. Wang, Y. ding, and J. Xu, "Design and implementation of a c/s structured distributed multi-thread crawler using ace and pvfs," Journal of Computational Information System, vol. 4, no. 3, pp. 883–890, August 2008.

[16] G. Feng, T.-Y. Liu, Y. Wang, Y. Bao, Z. Ma, X.-D. Zhang and W.-Y. Ma, "Aggregaterank: bringing order to web sites," in SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 2006, pp. 75–82.

[17] W. Jiang, Y. Guan, amd XL. Wang "A Pragmatic Chinese Word Segmentation System" Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, July 2006, pp. 189-192