Proceedings of the 10th WSEAS International Conference on CIRCUITS, Vouliagmeni, Athens, Greece, July 10-12, 2006 (pp158-162)

Feature Selecting Model in Automatic Text Categorization of Chinese Financial Industrial News

HUEY-MING LEE [1)] , PIN-JEN CHEN [1)] , TSUNG-YEN LEE [2)]

[1)] Department of Information Management, Chinese Culture University
55, Hwa-Kung Road, Yang-Ming-San, Taipei (11114), TAIWAN

[2)] Department of Securities, Bank of Overseas Chinese, Taiwan

Abstract: - This work focuses on selecting features in the automatic text categorization of Chinese industrial and financial news. We use feature selecting method for the characteristics of subclass Chinese financial and industrial news. However, it is an open challenge for subclass news in solving real-world problems which are often high-dimensional. Therefore, we proposed a feature selecting model in automatic text categorization of Chinese financial industrial news. This model can not only discover features from training news, but also can tune features through testing news. The proposed model help to classify subclass news, and it will be useful to knowledge management. Furthermore, feature selection has received considerable attention in improving the performance of the classification.

Key-Words: - Feature selecting; Text categorization; Chinese industrial and financial news

1 Introduction

With the prevalence of the Internet, users can easily retrieve the information what they want from Internet. Unfortunately, people could not promptly acquire knowledge from this information source without computer processing technology. Therefore, an efficient knowledge management becomes very important. In recent years, text categorization is recognized as one of the most important technologies for acquiring knowledge from huge amount of electronic documents. The goal of text categorization is the automatic assignment of documents to a fixed number of predefined categories. In general, each document can be in multiple, exactly one, or no category at all [8]. Therefore, it is an open challenge for subclass news in solving real-world problems which are often high-dimensional [12]. Furthermore, the number of recent researches on English text categorization is very large, nevertheless, little has been done for Chinese texts [7]. It is difficult to perform Chinese text classification with satisfactory performance due to its special language nature.

In this study, we proposed a feature selecting model in automatic text categorization of Chinese financial news. This model is extended work of our previous research on Chinese text mining [3,4,5,6]. We applied some of the text-mining technologies in data pre-processing to build this model. In order to transform ordinary text document into predefined database, an information extraction process (IE) is developed to extract necessary information from text documents, including natural language process technologies [4].In this study, we focus on selecting features in the automatic text categorization of Chinese industrial and financial news. Furthermore, we try to use feature selecting method for the characteristics of subclass Chinese financial and industrial news. We apply TFIDF to extract keywords. Furthermore, we hope to get the own characteristics of each subclass of Chinese financial industrial news. Gain ratio is the measure used to select the best attribute to be tested and represents how precisely the best by the attribute predict the distribution of classes [11]. We apply ratio gain to rank the features where at each subclass is located the features. After ranking features, we can update the features. This model can not only discover features from training news, but also can tune features through testing news.

The remainders of this paper are as follows: In Section 2, we introduce our selecting features model in the automatic text categorization of Chinese industrial and financial news. The global view of the proposed model is discussed. In Section 3, we discuss the feature selection algorithm Gain Ratio what we use in our model in detail. Section 4 concludes the paper and discusses future research.

2 Related work

In this section, we give a global view of the knowledge discovery model in Chinese documents. We applied some of the text-mining technologies in data pre-processing to build the feature selecting

model. This model is derived from the previous work from the knowledge discovery model in Chinese documents [3, 4, 5, 6]. As shown in Figure 1, the model is divided into two parts, saying pre-process and post-process. The pre-process takes the Chinese documents as input data. The pre-process includes Chinese segmentation and information extraction. The extracted information is stored in pre-defined format databases to represent the knowledge template. The post-process, Chinese Knowledge Discovery (CKD), is applied to a rule learner, named TextRise, to induce the knowledge templates into a set of rule base. Users discover interesting or helpful knowledge rules aided by a proposed interestingness measure from the rule set [3]. Therefore, users can easily obtain useful knowledge or information without having to read large text documents from the Internet or other sources.

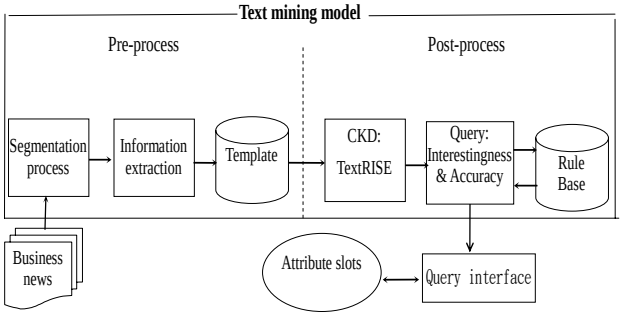

Fig. 1. The knowledge discovery model in Chinese Documents[4]

Unlike pre-process for English text documents, Chinese text documents are composed from Chinese characters without spaces. The process to divide Chinese text into segments, or phrases, is called segmentation process. There are three major approaches [1,2] to Chinese segmentation. The first approach is dictionary-based with maximum matching. That is, the process segments Chinese text by using a pre-defined Chinese dictionary. In general, the process takes the phrase with maximum length from all possible phrases. The second approach is based on statistical methodology. The model information is to divide Chinese text into proper phrases. The characters mutual information is statistical information derived from an existing corpus. The third approach integrates the first two approaches. We modified the third approach to the Chinese text segmentation process in this uses pre-produced characters database by mutual model.

Our previous work focuses on the post-process. In this study, we apply the pre-process to the feature selecting model in automatic text categorization of Chinese financial industrial news.

3 Feature selecting model in the text automatic categorization of Chinese financial industrial news

The feature selecting model in the text automatic categorization of Chinese financial industrial news is shown as in Figure 2. CFINITM can not only discover features from training news, but also can tune features through testing news. The proposed model help to classify subclass news, and it will be useful to knowledge management.

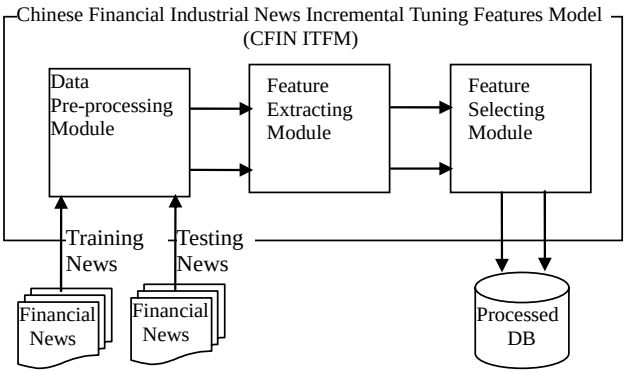

Fig. 2. Architecture of Chinese Financial Industrial News Incremental Tuning Features Model (CFINITFM)

There are three modules in CFINITFM, saying data pre-processing module (DPM), feature extracting module (FEM), and feature selecting module (FSM). In addition, there are three databases in the proposed model, saying segment word database (SWDB), keyword database (KDB) and processed database (PDB). The architecture of Chinese financial industrial news incremental tuning model is shown in Fig 2.

The functions of these modules are illustrated as follows:

- Data pre-processing module (DPM): We apply the pre-process technology of our previous work [3, 4, 5, 6], include segmentation etc. Pre-process of Chinese text documents are more difficult than English text documents. It can help us to deal with the Chinese text document through this module.
- Feature extracting module (FEM): This module derived from the previous work from the knowledge discovery model in Chinese documents [4]. We also use TFIDF to get the keywords to represent the features of the subclass.

- Feature selecting module (FSM): After getting the important keywords form FEM, we apply Gain-Ratio to rank the keywords and then select latest features through updating.
- Segment word database (SWDB): It stores the segment words which are generated from DPM.
- Keyword database (KDB): It stores the keywords which are extracted from FEM.
- Processed database (PDB): It stores the latest features which are tuned from FSM.

3.1 Data pre-processing module

Data pre-processing module (DPM) apply the pre-process technology of our previous work [3,4,5,6]. Pre-process of Chinese text documents are more difficult than English text documents. It can help us to deal with the Chinese text document through this module. Figure 3 depicts the integrated Chinese segmentation process [4] in the proposed model. We prepared mutual information from large corpus of Chinese characters. We also prepared a stop word list for removing meaningless characters. In the segmentation process, we use dictionary-bases and MI-bases to segment the same text. The segmentation process takes the longest phrase as a result.

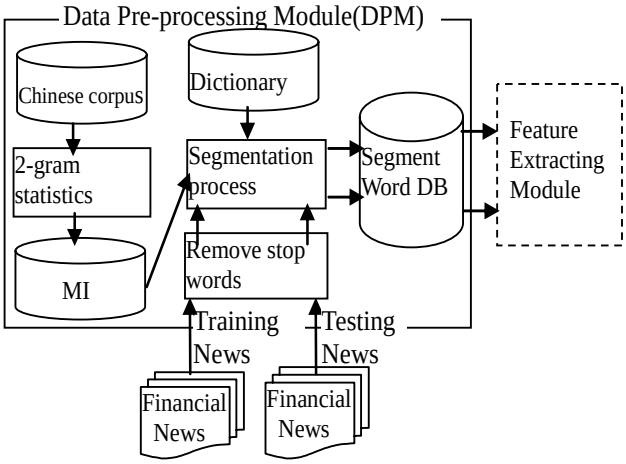

Fig. 3. Framework of data pre-processing Module

We use Sporat and Shih's approach [10] to calculate the Chinese mutual information (MI). The MI measure represents the concatenation strength of two Chinese character a and b. The MI value is calculated by the following equation [10]:

$$MI(ab) = \log_2(N) + \log_2\left(\frac{f(ab)}{f(a)\times f(b)}\right) \qquad (1)$$

where Chinese character b appears after character a. *f(ab)* represents the times that character b appears after character a. *f(a)* and *f(b)* represent the number of times that characters *a* and *b* appear, respectively. N is the total number of Chinese characters in the corpus. Chinese phrase ab could be a Chinese phrase if their MI value is high. Chinese character sequence abc could be highly possible a true Chinese phrase if both MI(ab) and MI(bc) are high. In this way, we can possibly find a new n-gram phrase. This approach solves the deficiency of phrase-based segmentation approaches for new phrases.

3.2 Feature extracting module

In the feature extracting module, we use TFIDF to get the keywords to represent the features of the subclass.

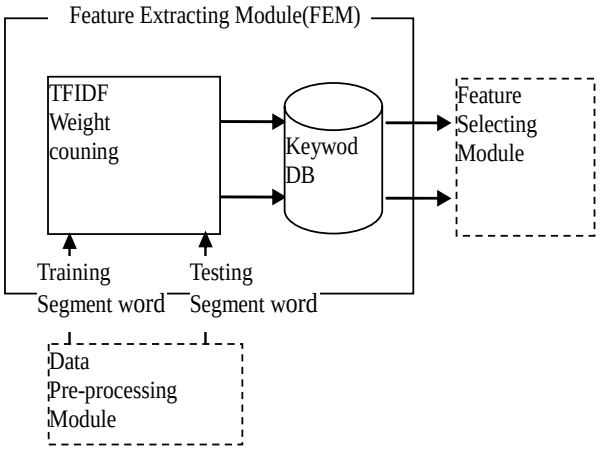

Fig. 4. Framework of Feature Extracting Module

Our work employed TFIDF method to compute the score of each term. We extract the several highest score terms as the keywords which are the type of subclass. We can select key terms by using TFIDF (term frequency; inverse document frequency) [9]. Term frequency is the number of times a particular term occurs in a given document or query. Inverse document frequency is a measure of how often a particular term appears across all of the documents in a collection. So, common words will have a low IDF and words unique to a document will have a high IDF.

The TFIDF weighting scheme is used to assign higher weights to distinguished terms in a document. TFIDF makes two assumptions about the importance of a term. First, the more a term appears in the document, the more important it is. Second, the more it appears through out the entire collection of documents, the less important it is since it does not characterize that particular document very well [9]. The weight for term t_i in a document d_i, W_i is defined as follows:

$$W_i = tf_i \times \log_2 \frac{N}{n} \qquad (2)$$

where tf_j is the frequency of term t_i in document d_i, N the total number of documents in the collection, and n the number of documents where term t_i occurs at least once.

3.3 Feature selecting module

After extracting features form FEM, we apply Gain-Ratio to rank the keywords which are the type of subclass and then tune the latest seleting features through testing news. The proposed feature selecting module (FSM) is shown as Figure 5.

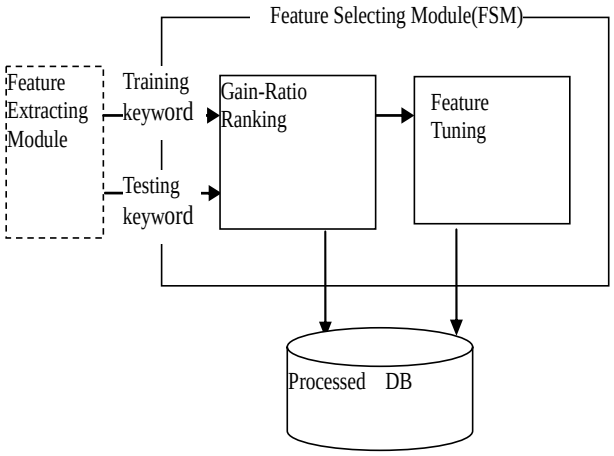

Fig. 5. Framework of feature selecting module

In this module, we proposed the utilization of the Gain-Ratio to get its own characteristics of each subclass Chinese financial industrial news.

We get the training keyword from feature extracting module and send to Gain-Ratio ranking these extracting features. After ranking the features, we can get its own characteristics of each subclass Chinese financial industrial news. Furthermore, we can not only discover features from training news, but also tune features through testing news. Through tuning strategy , selecting features can update the latest features. It will help to subclass of automatic text categorization, if we can get the latest feature.

4 Conclusion

With the prevalence of the Internet, users can easily retrieve the information what they want from Internet. Information explosion shows that efficient information summarization is aspired to all users. Therefore, an efficient knowledge management methodology becomes very important. Some technologies, such as automatic text categorization, for acquiring knowledge from huge amount of electronic documents are recognized as important technology in this field.

In this study, we proposed a feature selecting method for the characteristics of subclass Chinese financial and industrial news. This work focuses on feature selecting method for the characteristics of subclass Chinese financial and industrial news. However, it is an open challenge for subclass news in solving real-world problems which are often high-dimensional. Therefore, we proposed a feature selecting model in automatic text categorization of Chinese financial industrial news. This model can not only discover features from training news, but also can tune features through testing news. The proposed model help to classify subclass news, and it will be useful to knowledge management. Furthermore, feature selection has received considerable attention in improving the performance of the classification. It's a major contribution that our proposed model can process a mass amount of Chinese text documents and induce them into a knowledge management.

References:

[1] Hsu, C.-C., Chen, J.-K., Data Mining in Chinese News Articles, *Journal of Information Management*, 7(2), 2001, pp. 103-122.

[2] Hu, S.-J., Hsu, C.-C., Word Segmentation in Chinese News Articles, *Proceedings of the 10th International Conference on Information Management*, 1999, pp. 968-974, Taiwan.

[3] Huang, J.-Y., Lee H.-M., and Chen, W.-Y., Industrial News Knowledge Discovery with Text Mining Approach, *Proceedings of the 7th Conference on Information Management and Practice* (CSIM2001)[CD-ROM], 2001, Taipei, Taiwan.

[4] Huang, J.-Y., Lee, H.-M., Knowledg discovery model in Chinese industrial news, *Proceedings of the Second International Conference on Electronic Business* (ICEB-2002), 2002, pp.412-414, Taipei, Taiwan.

[5] Huang, J.-Y., Lee, H.-M., Automatic information extraction in Chinese industrial news, *Proceeding of the Third Conference on Information Management*, 2002, pp.861-869.

[6] Huang Ju-Yu, Lee Huey-Ming, Fang Chen-Liang., A Chinese Text Mining Application: An Automatic Answer Reply to Customers' E-mail Queries Model, *The Seventeenth International Conference on Software*

Engineering and Knowledge Engineering, 2005,Taipei, Taiwan.

[7] Lin Hung-Ru, Tsay Jyh-Jong ,Combining Classifier for Chinese Text categorization, Master thesis, *Department of Computer Science and Information Engineering National Chung Cheng University*, 2000, Chiayi, Taiwan.

[8] Mladenic D., Grobelnik M., Feature selection for unbalanced class distribution and naive Bayes. Machine Learning, *Proceedings of the Sixteenth International Conference*, 1999, pp. 258--267.

[9] Salton, G., Automatic text processing: The transformation, analysis and retrieval of information by computer, *Addsion-Wesley*, 1989, Massachusetts.

[10] Sporat, R., Shih, C., A statistical method for finding word boundaries in Chinese text, *Computer Processing of Chinese and Oriental Languages*, 1990, 4(4), pp. 336-351.

[11] Tatsunori Mori, Miwa Kikuchi, Kazufumi Yoshida, Term Weighting Method based on Information Gain Ratio for Summarizing Documents Retrieved by IR Systems, *Journal of Natural Language Processing*, 9(4), 2003,3-22.

[12] Yang Y., Liu X., A re-examination of text categorization methods, *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, 1999,pp.42-49.