A New Method of Text Feature Selection for Knowledge Discovery

Li Zhang[1], Xing Liu[2], Rong An[1], Xin Zhao[1], Kejia Yi[2, a]

[1]Systems Engineering Research Institute of CSSC, Beijing, China

[2]Science and Technology on Underwater Acoustic Antagonizing Laboratory, Beijing, China

[a]hearingme@yeah.net

Keywords: knowledge discovery; data mining; information model

Abstract. In the paper, we considered the problem to model the text document for knowledge discovery. It is urgent problem to find valuable information knowledge in the mass text documents. Data mining based on eigenvector is widely used to mine association between information of mass documents as much as possible for knowledge discovery. However, traditional information model methods for text are hard to achieve the feature vector for some special knowledge discovery task. This paper proposed a new feature selection method of text for knowledge discovery, which is useful to find the valuable words of the text for the special knowledge discovery. In addition, the dimension of the feature vector has been reduced by proposed method, which is of great help to improve the efficiency of data mining effectively.

Introduction

Now a day, the text document is spontaneously increasing over the internet, e-mail and web pages and they are stored in the electronic database format. It is urgent problem to find valuable information knowledge in the mass text documents. Knowledge discovery refers to take the sample from all the data collection, extracts and refines patters from data. Knowledge discovery tasks include data summarization, clustering, classification and deviation analysis. Knowledge discovery are widely used to mine association between information of mass documents as much as possible, linking relevant information really. Generally the knowledge discovery process can be roughly interpreted as Trilogy: data preparation, data mining and the interpretation and assessment of results. Data preparation is to select, preprocess and exchange data. Different algorithms can be chosen for the same mining task. Data mining is the core part of knowledge discover. Knowledge discovery methods include support vector machines, artificial neural network, fuzzy clustering, D-S evidence reasoning, Bays networks, rough set, fuzzy set, association rule, bionic optimization algorithm and so on. Association rules method is widely used to find valuable information knowledge in the mass text documents. The correlation or association is found from the text documents, which identifies property value set of higher occurrence frequency from data set, which is named as the frequent item sets. Information model is used for knowledge discovery from text documents using the association rules. There are two models used widely, model based on eigenvector and hierarchical model.

Mode text is to select text features, which refers to choose words group which can represent textual lexeme as textual features while categorizing texts. Chinese texts are composed of words, sentences and paragraph, which can easily be distinguished by these distinct marks, however, lexis of Chinese has no mark to be distinguished from. So it is necessary to segment words before selecting the textual features. Word segmentation refers to resetting the word serial according to a definite criterion [2].We called the serial as lexical entry. Nowadays, there are so many mature methods aimed at Chinese word segmentation, which can satisfy certain levels with a high accuracy rate [3][4]. After word segmentation, the text is transferred to many words. Then the frequency of each word is counted, which represents the feature item along with the words of the text. However, represented by feature item through word segmentation algorithm and word frequency statistics directly, the dimension of the vector shall be very large. If we adopt the untreated text vectors as the model of text, it will not only cost much time, to reduce the efficiency of the whole process, but also have little use for knowledge discovery. The results as above are not always so satisfied.

So we have to purify the text vectors on the base of maintaining its original meaning to find the text feature which can represent its text feature best. To solve this problem, the most effective way is to reduce dimension through feature extraction. The common way is to pick out the terms which make great contributions to knowledge discovery. The conventional selection strategy includes TF•IDF, DF,IG, MI, CHI, ECE and so on, which is of great help to reduce the dimension of the text vector. The text vector is finally composed of the words which exist frequently in the text. While the valuable words of the text for knowledge discovery with some special purpose such as military information discovery are not the most frequently existed in the text. This paper proposed a new text model method for military knowledge discovery.

Proposed text model method

Three categories of Chinese word segmentation have been proposed: dictionary-based methods, comprehension-based methods and statistical-based methods. It includes forward maximum matching, reverse maximum matching and shortest path in the dictionary-based methods. The input text is matched to dictionary which has been constructed to judge a string for a word if it is found in the dictionary or else. The advantage of the method is easily implementing. In this paper, the proposed text model method is based on dictionary. The proposed text model method composes of six steps: (1) three dictionaries should be built, the first one is the common dictionary widely used in Chinese word segment method, the second one is stop word dictionary, which is usually ready for use, and the third one is the special dictionary. The common and stop word dictionaries are usually read for use, and the special dictionary should be built according to the knowledge discovery purpose; (2) the input text is segmented using reverse maximum matching by the common dictionary, and the text is converted to the feature vector with high dimension; (3) the useless words in the feature is removed from the feature vector using the deleting stop words which is usually included in the stop word dictionary; (4) the dimension of the feature vector is reduced by match the feature vector to the special dictionary, and the valuable words have been chosen out, which have valuable information for the knowledge discovery;(5) weights of the words in the feature vector are calculated using the TF•IDF method;(6) the word, the weights of which are higher than the threshold, are chosen as the feature vector.

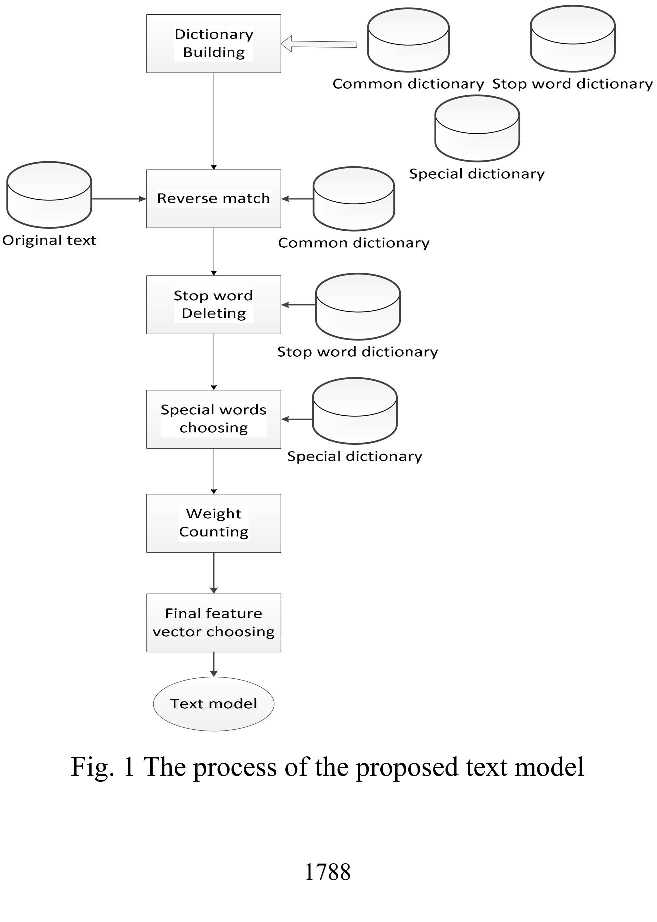

Fig. 1 The process of the proposed text model

Building Special Dictionary

Since they are widely used in the text processing, the common dictionary along with the stop word dictionary is usually already for use. Stop-words are words that from non-linguistic view do not carry information. Stop-words remove the non-information behavior words from the text documents and reduce noisy data. The key reason that interested knowledge as the military knowledge cannot be discovered is that the valuable words for the knowledge are not frequently existed in the document. If the special words are not chosen based on the special dictionary and the weight of all the words are counted, it results in that the words which are frequently occurred in the text will be chosen to build the feature vector, which has little interested knowledge. It is necessary to build a dictionary in advance. The special Dictionary can be dynamically built.

Reverse Match

It is shown in [5] that the error rate is 1 / 169 using simple forward maximum matching, and 1 / 245 using simply reverse maximum matching. Since accuracy of reverse matching is slightly higher than the forward matching and less ambiguity, reverse maximum matching is chosen in this paper. Reverse Maximum Matching (also known as RMM) ABC is a string of the text, if C∈W，BC∈W and ABC∉W（W is a word in the dictionary）, then it is segmented to word A and BC[6]. The reverse maximum matching method is used to segment words of the text based on the common dictionary.

Stop Word Deleting

Then the stop word dictionary is used to delete the stop words. Stop word dictionary is used that contains the words to be excluded. The Stop word dictionary is applied to remove terms that have a special meaning but do not discriminate for the knowledge discovery.

Special Word Choosing

The special word dictionary is used to choose the valuable words. Special dictionary is used that contains the valuable words that are useful for the knowledge discovery.

Weight Counting and Final Feature Vector Choosing

Weight counting is used to find the most frequently valuable words in the text. The metric weights all higher than the pre-specified threshold are selected as a key term.

TF•IDF is one of the most effective ways to calculate term weight. TF•IDF equals TF multiplies by IDF, in which TF is short for term frequency that is used to calculate the describing ability of the term; and IDF is short for inverse document frequency which is used to calculate the distinguishing ability of the term [7].

$$IDF = \log \frac{N}{n} \qquad (1)$$

where N is the text total of all categories, n is the number of the texts which include term t. The smaller n is, the larger the IDF is, and the term will have a better category distinguishing ability. If a word or phrase has a high TF value and hardly appears in other texts, it has a good category distinguishing ability and fits for classification and knowledge. We use the TF•IDF to describe the importance of the word. The words whose TF•IDF is higher than the pre-specified threshold are chosen to contributes the final feature vector.

Conclusion

Data mining based on eigenvector is widely used for knowledge discovery. Traditional information model methods for text are hard to achieve the valuable feature vector for some special knowledge discovery task. This paper proposed a new feature selection method of text for knowledge discovery, which is helpful to find the valuable words of the text for the special knowledge discovery. The proposed method reduces the dimension of the feature vector by many ways, which improves the algorithm efficiency effectively.

References

[1] James Allen. “Understanding the Natural Language,” (2nd edition). Translated by Liu Qun, etc. Beijing: The Electronic Industry Press, 2003

[2] Qu Wei-guang. “The Selection of Methodology on Automatic Words Segmentation of Chinese,” Computer Science, vol.29, 2002,pp. 54-56

[3] Sun Tie-li, Li Xiao-wei, Zhang Yan. “Automatic ChineseSegmentation Study in Information Filtering,” Computer Engineering& Science, vol.31, 2009, pp. 80-82

[4] Chen Ping,Liu Xiaoxia,Li Yajun, “Chinese word segmentation based on dictionary and statistics”, Computer Engineering and Applications. 2008 44(10).

[5] Zhang Xu. “A Chinese Words Segmentation Method Based on Dictionary and Statistics,” Cheng Du: The Electronic Science University, 2007

[6] Yubiao Dai, Xueli Ren,”A Hybrid Method to Segment Words,” ICALIP 2012.

[7] Qiaoyan Kuang, Xiaoming Xu, “Improvement and Application of TF-IDF Method Based on Text Classification”, Internet Technology and Applications, 2010 International Conference on , pp. 1-4.