A Study on the Classification of Dongba Literature

Yujing Chen[1,a], Ning Li[1,b], Xueqiang Lv[1,c]

[1] Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing 100101, China

[a]a1194743239@163.com, [b]ningli.ok@163.com, [c]lxq@bistu.edu.cn

Keywords: Classification,Dongba literature,Mutual information,Literature feature,SVM

Abstract. With dongba culture researches increasing year by year, there needs to be a highly efficient classification method to classify research achievements creating conditions for further study. Aiming at the shortcomings of the traditional mutual information method, giving full consideration to the factors such as word frequency, concentration and dispersion, and using the difference between the maximum and the second large value as a global evaluation function, GMI feature selection algorithm is proposed. Use this algorithm to choose text feature after one dimension reduction, and then get classification feature combined with the literature feature on secondary dimension reduction,and finally utilize the SVM to classify dongba literature. The experimental results show that the average accuracy rate and recall rate in all categories are 83% and 82% respectively. Experimental results show the proposed method is feasible in the dongba literature classification.

Introduction

Dongba culture is a precious cultural heritage of China and even the world. At present, the world cultural heritage has attracted attention of many scholars, the research achievements increased year by year. Dongba literature is the academic achievement which is published belonging to the periodical literature. The traditional classification method is based on artificial classification, but the method of artificial classification has many shortcomings. On the one hand, it need to invest a lot of manpower, material resources and energy; on the other hand, classification results has uncertainty and inconsistency. Aiming at this phenomenon, this paper tries to apply computer automatic classification technology to dongba literature classification, through the analysis of the characteristic of dongba literature, dongba literature feature selection methods are proposed and the method is utilized to extract classification feature, using SVM machine learning method to classify the dongba literature. Aiming at dongba literature, we get a kind of effective classification method.

Related Work

There are quite a lot of achievements for text categorization at home and abroad. However, dongba literature has not been researched in classification. Jinshu Su et al[1]introduced machine learning research progress in text classification and emphasized the use of support vector machine (SVM) in text classification. In the experiment[2], mutual information feature selection method is less effective. However, owing to its small computing time complexity and easytouse, it becomes one of the important feature selection algorithm[3]. Chi N W[4] proposed the text classification method based on ontology. The method applied ontology to match the text in the existing corpus, and carry out text classification. But the construction of ontology library workload is large, the efficiency is low.

A Feature Extraction Method for Dongba Literature

Text Feature Selection. The traditional mutual information (MI) calculation formula is shown in Eq.1[5].

$$\mathrm{MI}(\mathrm{w_i},\mathrm{c_j}) = \log\frac{\mathrm{p}(\mathrm{w_i},\mathrm{c_j})}{\mathrm{p}(\mathrm{w_i})\times\mathrm{p}(\mathrm{c_j})}. \quad (1)$$

In Eq.1 $\mathrm{p}(\mathrm{w_i},\mathrm{c_j})$indicatesthe probability of word $\mathrm{w_i}$appearing in the$\mathrm{c_j}$class,$\mathrm{p}(\mathrm{w_i})$ indicates the probability of the text which contains the item $\mathrm{w_i}$. $\mathrm{p}(\mathrm{c_j})$ indicatesthat the probability of the emergence of the document which belongs to the$\mathrm{c_j}$class.The total mutual information content of the item $\mathrm{w_i}$in the m categories is represented asEq.2.

$$\mathrm{MI}(\mathrm{w_i}) = \mathrm{p}(\mathrm{c_j})\sum_{j=1}^{m}\mathrm{MI}(\mathrm{w_i},\mathrm{c_j}).(2)$$

We can get the shortage of traditional mutual information[6][7]as follows.The traditional mutual information method only consider the document frequency without considering word frequency. From Eq.2, the traditional method is a kind of average ability.

We get the improved mutual information formula is shown in Eq.3.

$$\mathrm{GMI}(\mathrm{w_i},\mathrm{c_j}) = \partial\beta\gamma\mathrm{MI}(\mathrm{w_i},\mathrm{c_j}). \quad (3)$$

In view of the traditional method of mutual information without thinking about the word frequency,in this paper, we introduce the feature word frequency factor ∂. The formula is shown in Eq.4.

$$\partial = \frac{\sum_{i=1}^{n}\mathrm{TF_{ij}}}{\sum_{j=1}^{v}\sum_{i=1}^{n}\mathrm{TF_{ij}}}. \quad (4)$$

In the formula,$\mathrm{TF_{ij}}$is the frequency ofthe wordj in the text of i,n is the total number of text in this class, and v represents the total number of the words in this class.

This paper introduces the dispersion factor β.The dispersion factor β is the ratio of the number of text which contain the word and the total number of the class[6].

According to traditional problems without considering concentration degree in the mutual information, we introduce the concentration factor γ. The formula isshown in Eq.5.

$$\gamma = \frac{\mathrm{df}(\mathrm{w},\mathrm{c_i})}{\sum_{i=1}^{m}\mathrm{df}(\mathrm{w},\mathrm{c_i})-\mathrm{df}(\mathrm{w},\mathrm{c_i})+0.1}. \quad (5)$$

In the formula, $\mathrm{df}(\mathrm{w},\mathrm{c_i})$represents the number of text which contain word w in the class $\mathrm{c_i}$and m represents the total categories.

YufangZhang[8]proposed a new type of global assessment of function. We use the difference between maximum value and the second largest value as the evaluation function.we called it GMI feature selection algorithm.

Literature FeatureSelection. The title and abstract are very important for the literature.This paper chooses the title and abstract information as the literature feature. By text preprocessing, the title is divided into the lexical entry collection $\boldsymbol{s_1}$, the summary is divided into lexical entry collection$\boldsymbol{s_2}$.Literature feature is gotten by getting the intersection of$\boldsymbol{s_1}$and$\boldsymbol{s_2}$.

Featureselection methods for dongba literatureis as follows. For every word in the text feature set, if the term is also in the literature feature set, then the term was selected as the classification featureof dongba literature.

Feature Weighting and Document Representation. In this paper, the TF-IDF algorithm is used to calculate the contribution of the word to the text. In order to reduce the influence caused by the inconsistency of the length of the text, we adopt the normalized calculation formulaand we use the vector space model[9]converts each text to a set of vectors, then convert the text data into the form that the computer can process.

Experimental results and Analysis

Experimental Description. The experimental data of dongba literature is obtained by collecting from HowNet. Combined withCLC(Chinese Library Classification) and manual marking,we get the experimental data. The number of the total data is 983, and randomly divided into training data and testingdata according to the ratio of 4:1. It is divided into six categories as shown in Table 1. We use the correct rate, recall rate and F-Measure[10]to evaluate the dongba literature classification.

Table 1 The distribution on the all kinds of dongba literature

Category	Categoryfor shor	Total number	Training data	Testing data
Tourism and Customs	F	91	72	19
Cultural and Science	G	179	143	36
Literature and Art	I	168	134	34
Languages	H	261	208	53
Philosophy and Law	B	181	144	37
Other	O	103	82	21
Total number		983	783	200

Experimental Results and Experimental Analysis. The article adopts the method of MI, GMI, literture feature, MI is combined with literature feature and GMI is combined with literature feature, and feature a, b, c, d and e is obtained respectively. After feature weighting and document representation, we use SVM for training and testing. Fig.1 is the change chart of the correct recognition rate along with the change of the dimension of the text feature under the five kinds of features.

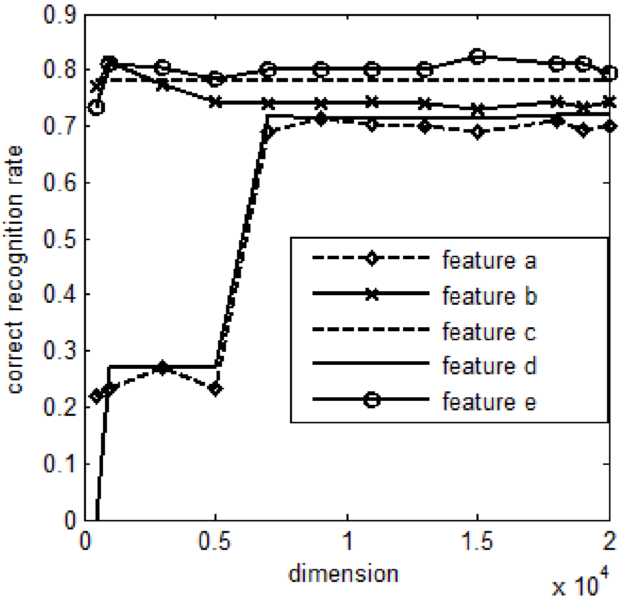

Fig.1 The change chart of the correct recognition rate along with the change of the dimension of the text feature under the five kinds of features

As can be seen from Fig.1, using the GMI feature selection method, the correct recognition rate is higher than the MI feature selection method in each dimension. Generally speaking, both GMI method or the MI method, correct recognition rate is higher than original method through combining with the literature feature. This illustrates the literature feature on the importance of dongba literature classification. However, only using the literature feature as feature of dongba literature classification, the recognition correct rate is 78%. From the above chart we can find that the MI method first reduce dimension to get the text feature, and then combine with the literature feature of second dimension reduction method, and correct recognition rate is higher than byonly using the literature feature as the classificationfeature, indicating that two dimensionality reduction can choose more precise classification feature,so as to improve the correct recognition rate of the test set. In this paper, the method is tested on 200 corpora,from Fig.1, we can see that this method has the best effect in the text feature selection 15000 dimensionand 634 dimension classification feature is obtained by combining with the literature feature, and the 165 is correctly identified. Average correct rate and recall rate were 0.83 and 0.82.

Conclusion

In this paper, the classification technology is introduced into the dongba literature category. After analyzing the characteristic of dongba literature, suitable for dongba literature classification method is proposed. Finally, we use SVM classifier for training and testing the dongba literature. Through the experiment, it is proved that the average accuracy rate and recall rate of the method in this paper are more impressive.

Acknowledgement

This paper is sponsored by Seminar of National Scocial Funds Project (GN: 12&ZD234).

References

[1] Jinshu Su, Bofeng Zhang, Xin Xu. Advance in machine learning based text categorization [J]. Journal of softare,2006,17(9):1848-1859. In Chinese.

[2] Yang Y, PedersenJ O.A comparative study on feature selection in text categorization[C]//Proceedings of the 14th International Conference on Machine Learning(ICML297),1997:412-420.

[3] Bakus J, Kamel M S. Higher Order Feature Selection for TextClassification[J]. Knowledge and Information Systems, 2006,9(4): 468-491.

[4] Chi N W, Lin K Y, Hsieh S H. Using ontology-based text classification to assist Job Hazard Analysis[J]. Advanced Engineering Informatics, 2014, 28(4): 381-394.

[5] Xiaoli Fan, Xiaoxia Liu. Study on mutual information-based feature selection in text categorization [J]. Computer Engineering and Applications,2010,46(34):123-125. In Chinese.

[6] Kai Lu . Improvement on mutual information in feature selection [D].Wuhan: Central China Normal University,2014. In Chinese.

[7] Caifeng Zheng . Study of mutual information feature selection in Chinese text classification [D]. Chongqing: Southwest University,2011. In Chinese.

[8] Yufang Zhang , Binhou Wan,Zhongyang Xiong. Research on feature dimension reduction in text classification[J].Application Researchof Computers,2012,29(7):2541-2543. In Chinese

[9] Chengqing Zong. Statistical natural language processing [M].The second edition Beijing: Tsinghua University Press, 2013：417-419. In Chinese.

[10] Jiana Meng, Hongfei Lin, Yangpeng Li. Application of feature selection method to text categorization based on feature contribution degree[J]. Journal of Dalian University of Technology, 2011, 51(4): 611-615. In Chinese.