A Part of Speech Based Public Opinion Text Classification Method

Rui Liu[1,a], Zhiqiang Wei[1,b] , Hao Liu[1,c*] ,Qianqian Fu[1,d]

[1] College of Information Science and Engineering, Ocean University of China,Qingdao,Shandong,266100,China

[a]liurui19911116@126.com, [b]weizhiqiang@ouc.edu.cn, [c]liu.hao@ouc.edu.cn, [d]fqq0902@163.com

Keywords: Public opinion; Text categorization; Part of speech; Feature extraction

Abstract. An improved text classification algorithm is presented to improve the accuracy and efficiency of the public opinion classification. The algorithm filters the part of speech before feature extraction to decrease the useless feature and then classifies text according to the calculated weight. The experimental results show that the feature extraction of the improved algorithm is more efficient than the previous ones, and the text classification results in different feature dimensions are more accurate, especially in the lower dimensions. Therefore, it has important significance for text classification by analyzing the weight of the part of speech to extract feature and calculate weight before classification.

Introduction

E-government construction is the key project of information technology construction in China [1]. However, with the development of Internet information technology, the e-government shouldn't stop at the office automation phase in order to meet the needs of modern development. In recent years, the intelligent government has become a new model of governance and social development, more and more people are willing to participate in government construction through interaction between government and the public, and they are accustomed to express their demands and views on the internet. According to the statistic of a sub-provincial city's Mayor-mailbox in China, more than forty thousand e-mails and two hundred thousand phone calls were received, and CNNIC also shows that, the internet users in China reached 668 million in 2015. So the information range of e-government is much wider and the number of feature becomes much more than before. Although the methods based on statistical theory and machine learning have already been applied to the information automatic processing [2], however, the number of characteristics is too large to deal with no matter which feature selection method is used, and these methods can't solve many problems, for example, the wide source and huge content of information, colloquial, and the interference information. Therefore, combined with the characteristics of the part of speech, we calculate the characteristics and weights of all the parts of speech to filter the useless feature before feature extraction. The experimental results show that the improved algorithm is more accurate and efficient.

The rest of this paper is organized as follows. Section II presents the key technologies of text classification analysis algorithms, section III describes the feature extraction methods based on part of speech, section IV describes the system related functions implementation, and section V shows the experimental results and analysis. Conclusion and the future work are presented in the last section.

Related Works

Recent years, many kinds of feature selection methods have been introduced and widely used in text categorization [3], among them document frequency, information gain and chi-square statistic are the commonly used feature selection methods now.

Document Frequency (DF) is a simple and effective feature selection method which computes the frequency of every feature in the text [4]. The small DF of a feature indicates that it is a low frequency word which has no representation of category. On the contrary, the large DF of a feature appeared in different categories indicates that the feature has little contribution to text categorization. DF has the advantage that it has good effect in practice with small amount of calculation. But the disadvantage of DF is, the rare feature shown in the specific class, which could reflect the characteristic of the class, would be filtered because it is lower than the threshold, so it has great influence on the classification accuracy.

Information Gain (IG) is frequently used as a criterion in machine learning [5]. The measure standard of feature's Information Gain is that the larger information the feature owns, the more important the feature is for categorization. The disadvantage of information gain is that it takes the un-happened condition into account. Though the un-happened condition could have good effect on the text classification, the interference of it has greater influence than the good effect.

χ^2-statistic (Chi-square) is used to measure the correlation degree between feature words and the texts category [6], which assumes there is a χ^2 distribution with first order freedom between them. The greater the χ^2 value of the feature word for one category is, the more correlation between them is and the more category information the feature word has, therefore the less independence of the feature word is.

Feature Extraction Methods Based on Part of Speech

Those feature extraction methods described above have their advantages in English sample sets, but they have no efficiency in Chinese sample sets. There are two main reasons, one is that feature extraction needs large amount of calculation and the efficiency is very low, the other one is that the dimension of feature vector is too high.

We have analyzed some feature sets obtained by traditional feature selection methods and found that each feature set has many pronouns, quantifiers and character strings, which accounted for about 4% to 8% of the total feature we extracted. According to the study of Han [7], the noun is the most important part of speech to represent text content and its classification results are similar to the classification results by using all features we extracted. However, it is more effective by applying the combination feature of the nouns, verbs, adjectives and adverbs to cluster, which is more accurate and efficient than the combination of all the features. So, to reduce the feature number, the algorithm presented in this paper selects the combination of nouns, verbs, adjectives and adverbs as feature and extracts the feature after speech filtration in the segmentation step. The feasibility and efficiency of the algorithm would be verified by experiment. In order to tagging the part of speech, we choose the Institute of Computing Technology Part of Speech Set as our part of speech tagging sets, which has 99 tagged parts of speech, to satisfy the experimental precision.

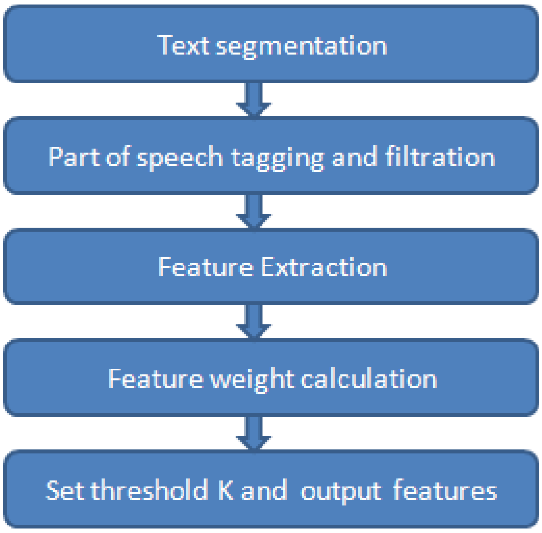
Text segmentation
Part of speech tagging and filtration
Feature Extraction
Feature weight calculation
Set threshold K and output features

Fig.1. The process of the feature extraction method based on the part of speech

We adopt the feature extraction method based on the part of speech and regard the nouns, verbs, adjectives extracted in the Chinese texts as the first level characteristic words, and then do feature selection of these words. Then we sort and select the top K characteristic words according to the weights computed by the characteristic frequency and document frequency for experiment. The process of the feature extraction method based on the part of speech is shown in Fig.1.

System Related Functions Implementation

A. Feature weight calculation

Term Frequency-Inverse Document Frequency (TFIDF) is one of the weight calculation methods commonly used in information retrieval and data mining [8]. The main idea is that a word or phrase has better category classification ability if it has higher appearance rate in one specific article than in other articles. TF is the word frequency, which is used to denote the frequency of characteristic t in the text d. IDF is the inverse document frequency which is the quantization of the distribution of characteristic in the text space. TF-IDF combines TF with IDF to measure the related properties of one characteristic in the text space. As one of the most commonly used weighting methods, TF-IDF has advantages of excellent accuracy and efficiency. However, the position information of words cannot be reflected in TF-IDF, so that there is no difference of the computing method of feature between the words in the title and the text.

B. Text representation model

Value Stream Mapping (VSM) is proposed by Professor Salton in 1968 and applied to the famous SMART text retrieval system successfully [9], it has become one of the most convenient and effective textual representation models. The starting point of VSM is that each text is composed by the feature which can reveal its content and be independent to other feature, and each feature could be regard as a dimension of vector space so that the text could be expressed as the collection of feature so as to ignore the complex relationship between the paragraphs and sentences in the text. Therefore, the text can be represented by a vector and the similarity between texts can be calculated by the distance between their vectors. However, VSM has the premise that each feature should be independent to others, so it is not suitable for feature combination.

C. Performance Measures

To evaluate the classification effect [10], we use the common performance evaluation methods: the Recall rate R, the Precision P and F1. The recall rate is the correctly probability of identified sample in a category, which is defined as the ratio of the correctly classified texts and the total texts. The precision is the correct probability of the classifier, which is defined as the ratio of the correct texts and the total texts classified by the classifier into the category. Generally, the single measurement F1, which is combined by the recall rate and the precision rate, is often used to compare. The formula of F1 is shown in Eq.1.

$$F1 = 2RP / \left(R + P\right) \tag{1}$$

Experimental Setup and Result

A. Data Sets

We use the letters from Qingdao Mayor-mailbox between 2013 and 2014 as the Corpus in this experiment. There are six categories and each category contains 400 randomly selected texts, of which 300 texts would be used as the training set and the remaining 100 texts would be used as test set. So there are 1800 training texts and 600 test texts would be used.

B. The part of speech tagging

Before the classification, we need to set word segmentation and part-of-speech tagging for the training set. The results of the part-of-speech tagging are shown in Fig.2. We can see from the statistics of the training set that the nouns and verbs have the highest frequency. And in addition to the nouns, verbs, adjectives and adverbs, there are about one fifth remaining words.

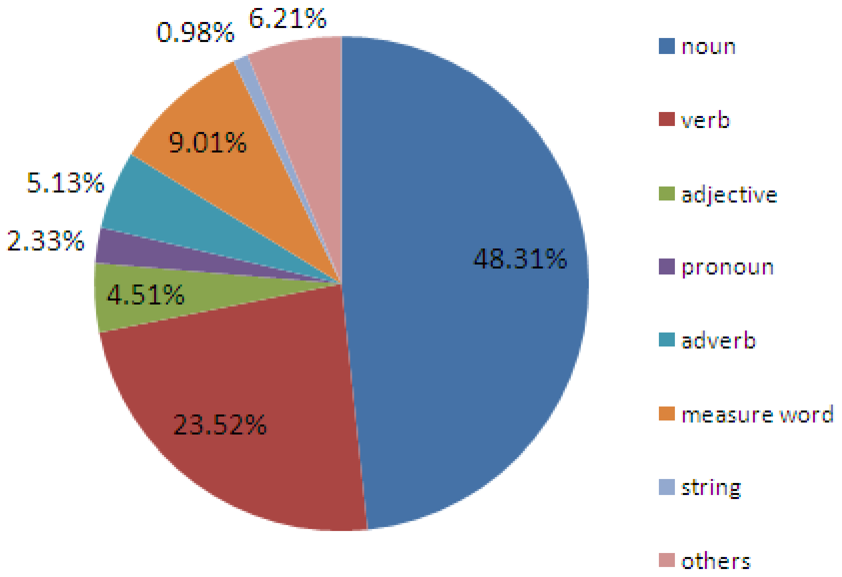

Fig.2. Proportion of words

C. Feature selection and weight calculation

We choose the traditional feature selection methods, such as the Information Gain, Document Frequency and Chi-square Statistic, to extract feature in and without use of part of speech respectively. We will test every 200 increase of feature from the beginning 200 by changing the threshold size K and compute the weights by TF-IDF method, then the experiment will learn by SVM and the classification accuracy of the algorithm would be tested. Then we calculate each group's recall rate and precision rate, and use Eq.1 to get the F1 value of each feature extraction methods.

D. Experiment results and analysis

Fig. 3 shows the mean value curves of F1 of the three different feature extraction methods tested by libsvm with the part of speech filtration being used or not respectively. The horizontal axis is the number of features; the vertical axis is the value of F1. The curves tagged with POS (part of speech) show the feature extraction results of filtering the part of speech.

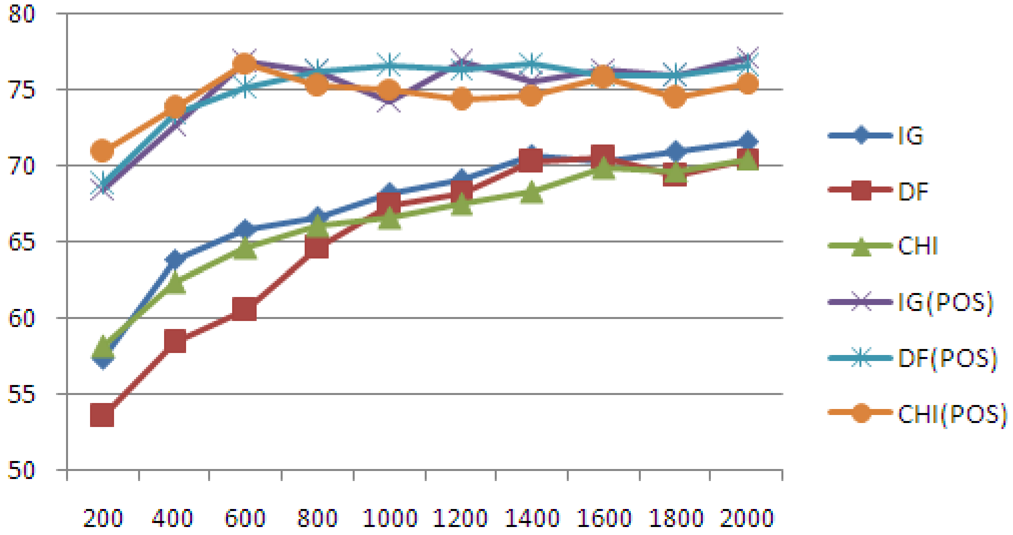

Fig.3. F1 values of each feature extraction methods

We can see that when the feature dimension is the same, the accuracies of the 3 feature selection methods with the part-of-speech filtration have been promoted obviously to 75%, especially in the lower feature dimension. The experiment results show that the feature extraction with the part-of-speech could improve the accuracy rate of government public opinion text classification.

Conclusion

The public sentiment classification method based on part of speech meets the needs of public opinion classification by applying the part of speech filtration to the feature extraction process. Firstly, we extract the composition of feature set by analyzing the traditional feature set and get the theoretical feasibility of applying the part of speech to the feature extraction. Secondly, we verify the feasibility of text feature extraction based on the part of speech by experiment, and the accuracy and the computational efficiency are improved than the traditional feature extraction method. However, the algorithm presented in this paper also can be modified by applying it to genetic text classification and dividing the characteristics of part-of-speech, for example, the noun can also be divided into personal names, place names, azimuth noun, proper noun, and so on. We can also filter the feature by the part-of-speech to future improve the classification results.

Acknowledgments

This research is supported by Qingdao People's Livelihood Science and technology program (14-2-3-62-nsh), and Shandong Province Science and Technology Development Plan (2014GGX101005).

References

[1] Layne K, Lee J. Developing fully functional E-government: A four stage model[J]. Government Information Quarterly, 2001, 18(2):122-136.

[2] Sebastiani F. Machine learning in automated text categorization [J]. Acm Computing Surveys, 2002, 34(2):1-47.

[3] Aas K, Eikvil L. Text Categorisation: A Survey. [J]. Raport Nr, 1999.

[4] Sanderson M, Joho H. Document frequency and term specificity[J]. In Recherche d'Information Assistée par Ordinateur Conference (RIAO), 2007, 16(1):321-342.

[5] Lee C, Lee G G. Information gain and divergence-based feature selection for machine learning-based text categorization [J]. Information Processing & Management, 2006, 42(1):155-165.

[6] Zheng C, Xiong D K, Liu Q Q. The Short Text Classification Method Based on CHI Feature Selection and LDA Topic Model[J]. Computer Knowledge & Technology, 2014.

[7]Pu Han, Dongbo Wang, Yanyun Liu et al. Influence of Part-of-Speech on Chinese and English Document Clustering [J]. Chinese Information Technology, 2013, 27(2):65-73.

[8]Congying Shi, Chaojun Xu, Xiaojiang Yang. A Survey of TFIDF arithmetic. Computer Applications, 2009, 29:167-170.

[9] Salton, G, Wong, A, Yang, C. S. A vector space model for automatic indexing[M]. Morgan Kaufmann Publishers Inc, 1997, 273-280.

[10] Cheng Z, Lin S. Methods on Accuracy Evaluation of Text Classifier[J]. Journal of the China Society for Scientific & Technical Information, 2004.