TELKOMNIKA, Vol.15, No.1, March 2017, pp. 383~390
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v15i1.4303

A Novel Text Classification Method Using Comprehensive Feature Weight

Jian Liu*[1], Weisheng Wu[2]
[1]Information and Computer Engineering，Pingxiang University, Pingxiang, Jiangxi 337055, P. R. China
[2]Modern Educational Technology Center, Pingxiang University, Pingxiang, Jiangxi 337055, P. R. China
*Corresponding author, e-mail: liujianpxu@163.com

Abstract

Currently, since the categorical distribution of short text corpus is not balanced, it is difficult to obtain accurate classification results for long text classification. To solve this problem, this paper proposes a novel method of short text classification using comprehensive feature weights. This method takes into account the situation of the samples in the positive and negative categories, as well as the category correlation of words, so as to improve the existing feature weight calculation method and obtain a new method of calculating the comprehensive feature weight. The experimental result shows that the proposed method is significantly higher than other feature-weight methods in the micro and macro average value, which shows that this method can greatly improve the accuracy and recall rate of short text classification.

Keywords*: text classification, text categorization, comprehensive feature weight, feature selection*

1. Introduction

In recent years, with the development of instant communication technology and the popularity of internet-based applications, QQ chat, Facebook, micro-blog, news comments and other short text data is growing exponentially, and has become an important way of information transmission. The emergence of a large number of short texts has both advantages and disadvantages. On the one hand, the short text classification has important commercial value and application background in question answering community, advertising classification and so on. On the other hand, it is also an important carrier of harmful network information such as the bad public opinion and the rubbish information. Therefore, short texts classification, short texts clustering, topic detection, and tracking of short texts have become a focus in the field of data mining and information security [1-3].

The short text has generally only about 100 words, the text length is short, the noise data is more, the useful information contained is very little, so the information can be extracted is deficient [4,5]. Therefore, it put forward higher requirements to the related research of short text classification. Although the existing text classification field has made great achievements, it is difficult to apply directly to the short text. Short text classification is faced with two major problems in the field of text classification: high vector dimension and sparse feature [6-9], which resulting in unbalanced of the categorical distribution of short text corpus.

So far, there are two ways to improve the performance of short text classification. One is to use the external data source or knowledge base (such as WordNet, Wikipedia, etc.) to expand the feature of short text [10-12]. This method is simple and intuitive, can solve the problem of sparse feature words, and can achieve good results, but the classification accuracy needs to be improved. The other methods [13-17] are to improve the classification precision and recall rate by improving the performance of the classifier, but related articles to solve this problem from the point of view of the short text feature weight calculation is very few. Sriram et al [18] improved the classification precision of short text by combining the improved short text feature weight formula and feature extended algorithm based on the collinear relationship of features. In some articles [19-21], it is used in another method, which uses implicit semantic analysis to reduce the dimension of short text and noise.

Then we use the independent component analysis method to extract the most expressive features. These methods are combined with other methods to improve the

Received November 15, 2016; Revised January 10, 2017; Accepted January 30, 2017

effectiveness of short text classification, which cannot explain the effectiveness of the use of feature extraction method alone. Furthermore, we take into account the situation of the samples in the positive and negative categories, as well as the category correlation of words, so as to improve the existing feature weight calculation method and obtain a new method of calculating the comprehensive feature weight.

2. Feature Weight of Text

Explaining research chronological, including research design, research procedure (in the form of algorithms, Pseudocode or other), how to test and data acquisition. The description of the course of research should be supported references, so the explanation can be accepted scientifically.

In order to convert the text into a form that a computer can understand, a text is represented as a vector to be analyzed and calculated. Vectoring is the basis of text processing. Given a certain weight to the words in the text which mean the importance of a word to the characterization of the text, the greater the weight, the more important word to the text.

In the information retrieval field, the term weight calculation method is divided into two categories: one is the unsupervised *tf* (term frequency) and $tf \times idf$ (term frequency-inverse document frequency) method; another is the supervised method like $tf \times ig$ ($tf \times information$ gain), $tf \times gr$ $tf \times gain$ (ratio) and so on. The term weight calculation method mainly references the feature selection algorithm, as feature selection. These terms are also given different values to measure their contribution to text classification. However, these term weight calculation methods are mainly for the long text, and cannot be directly applied to the short text.

This is due to a prominent feature of the short text that the sample distribution is very uneven, that is, the number of samples in the dataset is much larger than other categories, resulting that small category text is submerged in a large number of other types of documents and it is difficult to identify. But in the massive text data to be processed, sometimes the data that system really cares about is only a small part, for example, in the network public opinion analysis and the hot sensitive topic detection and tracking problem, the valuable data in the practical environment is a small proportion. We refer the category has little samples as the positive category and the category has more samples as the negative category.

However, the existing feature weight calculation methods treat all categories in the same way, in fact, when applying the traditional method of long text feature weight calculation on the short text, the result of text classification is more inclined to negative category while ignoring normal category. The problem is particularly evident in the short text classification performance, leading to no high short text classification precision and recall rate, which cannot meet the need of practical application.

3. Comprehensive Category Feature Selection Method

In this paper, we regard the current category of short text dataset as a PC (Positive Category). In addition to the current category, another category is regarded as the NC (Negative Category). Related element information is shown in Table 1.

Table 1. Data Element Table

Feature	The number of samples including /not including *s* in *PC*	The number of samples including /not including *s* in *NC*
s	*sP*	*SN*
$\overline{s}$	$\overline{s}p$	$\overline{s}N$

s and $\overline{s}$ represent the circumstance of feature term *s* appear and disappear in sample respectively. *sP* represents the frequency of samples which contain feature *s* in *PC*; $\overline{s}p$ represents the frequency of samples which do not contain feature s in *PC*. *SN* represents the

frequency of samples which contain feature *s* in *NC*; $\overline{s}N$ represents the frequency of samples which do not contain feature *s* in *NC*.

Based on the idea of plain text, this paper has 3 important related analyses: 1) For a given term that contains s_i, if it occurs frequently in the text, that is when the document frequency is high, it will have a strong expression ability; 2) When the frequency of s_i in positive category is higher than that in the negative category, it shows that it has good classification ability which is called inverse document frequency; 3) When the ratio between the frequency of s_i not in negative category and the frequency of s_i in positive category is high, this shows that the relevancy frequency between s_i and the text category is high.

According to the above analysis, this paper proposes a feature selection method for short text. Since this method takes into account the case of the sample in the positive and negative categories, it is named as the integrate category IC (Integrate Category), the calculation formula is as follows:

$$IC(s_i,c) = df \times rf \times icf \tag{1}$$

where *df* is the document frequency, for a given term s_i, the value of *df* is derived from the calculation $\log(sP+1)$, and it represents the document frequency of term s_i in the positive category; *tP* is the number of samples which contains s_i in positive category. *rf* is the relevance frequency derived from the calculation $rf = \frac{\overline{s}N}{sP+1}$. *rf* is proportional to the frequency of *s* not in the negative category and inversely proportional to the frequency of *s* in the positive category. The bigger the *rf* value, the more the correlation between the term *s* and the category. *icf* is inverse document frequency and is derived from the calculation $icf = \log(\frac{|C|}{cf}+1)$, |C| is the total categories, c_f is the number of categories which contains term s_i.

In the case of a certain number of positive samples, the sample distribution of dataset can be roughly divided into the following three conditions: 1) the number of negative samples is more than the positive, and it is more evenly distributed in different categories; 2) the number of negative samples is less than positive, and it is more evenly distributed in different categories; 3) the number of negative samples is less than the positive, and the distribution is not balanced, that is, in a small number of categories. In order to express the above relations more visually, 20 categories are given, including 3 simple term examples of 200 short texts, simulating the distribution of positive and negative samples in the short text dataset, see Table 2.

Table 2. Comparison of the Relationship between PC and NC Terms

Term	$\frac{NC}{C2}$	$\frac{NC}{C3}$	$\frac{PC}{C1}$	sp	cf	$df \times icf$	$IC(df \times icf \times rf)$
s_1	20	100	0	20	3	8.538	39.294
s_2	20	4	4	20	4	7.636	68.478
s_3	20	80	80	20	4	7.636	15.526

As can be seen from table 2, the distribution of samples contain s_2 and samples contain s_3 in the positive and negative categories are very different, among which samples contains s_3 are mostly in the negative class, while the proportion of samples contains s_2 in the negative class is small, but the $df \times icf$ value of the two are the same; the value of s_2 is significantly increased after taking the relevance frequency of *s* in the positive category and negative category into consideration. In contrast, the increase in the value of s_3 is smaller, so that the gap between s_2 and s_3 is increasing. The frequency of s_1 in positive category is similar to that of s_2 and s_3, but its distribution range in negative category is small, it is concentrated in one category,

so the $df \times icf$ value of s_1 is slightly larger than that of s_3, and the $df \times icf \times rf$ value is between s_2 and s_3, this is consistent with the previous analysis.

The global feature weight of feature s for the entire corpus is shown in formula (2):

$$IC_{\max}(s) = \max_{1 \le j \le m} \{IC(s, C_j)\} \quad (2)$$

The advantage of this method is that it takes into account the distribution of entries in a single sample, and takes into account the category information of the text. The relevance evaluation of the terms in the positive and negative categories makes the feature words more discriminative in the imbalanced datasets. At the same time, selecting the maximum value of the global feature can help us obtain the feature with the highest category discrimination degree.

Algorithm description is as follows:

1) To pre-process the documents in the training corpus: word segmentation, remove the stop words;
2) Calculate $IC(s, C_j)$ of each feature term and category;
3) Calculate $IC_{max}(s)$ of all categories based on the second step results;
4) Sort *IC* value in descending order, take the first *M* values as feature words to be retained, *M* is the dimension of feature space. In this paper, the *M* value is 820.

4. Results and Analysis

In order to verify the performance of the short text feature weight algorithm, K-Nearest Neighbor (KNN) classification algorithm is adopted to classify the short text. KNN is a traditional pattern recognition algorithm, the algorithm is simple and intuitive, the classification accuracy is high, and the new training text does not need to be trained, so as to reduce the training time, it is widely used in text automatic classification. *K* represents the number of nearest neighbor samples in a class. Testing different odd value of *K* in the range of [3, 35], and use the result of the optimal *K* value for comparison.

4.1. Experimental Design

The multi classification problem can be decomposed into two classification problems, so the focus of this paper is to improve the classification performance of positive and negative two categories. At present, there is no general short text dataset, so this paper is based on the API of sina micro-blog to grab the micro-blog data as the text corpus. We select 8478 micro-blog s, each micro-blog has more than 6 words and the average text length is 42. We take about 100 micro-blogs of the Chinese super league in 2015 as a positive category, and 4236 micro-blogs of the traffic jam as a negative category.

Since the sample of dataset is very unevenly distributed, we use a 5-fold cross-validation method to carry out the classification results comparison. That is split the dataset into 5 folds, train on 4 folds and evaluate on the 5th, then iterate each fold as the validation fold, evaluate the performance and finally average the performance to avoid the contingency of experimental results and make sure the training data and testing data have no intersection. The results thus obtained can be considered to be credible.

4.2. Evaluation Method

There are three currently used classification performance evaluation index: precision, recall, and average test value. For short text classification with uneven distribution of samples, precision and recall ration will ignore the influence of small category PC. Micro-average and macro-average are two methods for global evaluation of classification results: micro-average is the arithmetic average of the performance indicator for each instance document; macro-average is first to calculate the classification results of each category then calculate the average value of all categories. Specific definitions are as follows:

$$MicroA = 2 \times \frac{MicroP \times MicroR}{MicroP + MicroR} \quad (3)$$

$$MacroA = 2 \times \frac{MacroP \times MacroR}{MacroP + MacroR} \tag{4}$$

where *MicroP* and *MicroR* represent the precision and recall rate of micro-average respectively; *MacroP* and *MacroR* represent the precision and recall rate of macro-average respectively. Before calculating the evaluation index, first we introduce the data elements that are related to the evaluation index, as shown in Table 3.

Table 3. Corresponding Table of Data Element

Retrieval	Relevant	Irrelevant
Yes	*RY*	*NR*
No	*RN*	*NN*

Suppose the total number of text is 1, then:

$$MicroP = \frac{\sum_{l=1}^{N} RY_l}{\sum_{l=1}^{N} RY_l + \sum_{l=1}^{N} NR_l} \tag{5}$$

$$MicroR = \frac{\sum_{l=1}^{N} RY_l}{\sum_{l=1}^{N} RY_l + \sum_{l=1}^{N} NR_l} \tag{6}$$

$$MacroP = \frac{1}{N} \sum_{l=1}^{N} P_l \tag{7}$$

$$MacroR = \frac{1}{N} \sum_{l=1}^{N} R_l \tag{8}$$

Where

$$P_l = \frac{RY_l}{RY_l + NR_l}, R_l = \frac{RY_l}{RY_l + RN_l}.$$

4.3. Experimental Results and Analysis

In practical application, the proportion of positive category text is very small. In order to compare the effect of negative category text data size on different feature weight method, let the initial positive and negative micro-blog text have 100 each, gradually increase the number of negative text, until the total text number reached 4300. Under the KNN classifier, the *MicroA* values and *MacroA* values of 6 different feature weight calculation methods are shown in Figure 1 and Figure 2 respectively. As can be seen, with the increase of data size, the performance value of all the feature weight method is on the rise. The reason may be that in a certain number of positive categories, when the data size is small, the text of each category is not sufficient to characterize this category, resulting in lower classification accuracy. This situation improves with the increase of data size.

The result of the experiments on Figure 1 and Figure 2 shows that since the *MacroA* is the average of each category and is greatly influenced by small categories, the *MacroA* values of different feature weight calculation method is greater than the *MicroA* values generally.

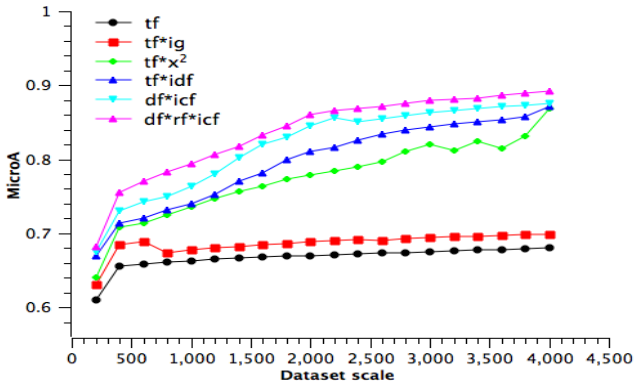

Figure 1. *MicroA* Value of Different Features Weight Calculation Method

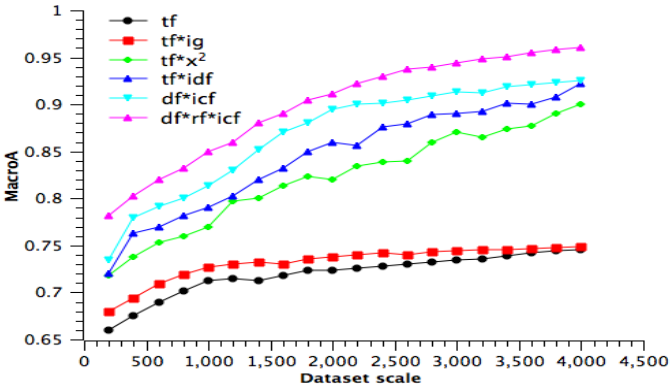

Figure 2. *MacroA* Value of Different Features Weight Calculation Method

Figure 3 shows the average of the highest classification result value of 5 times repeated trials on 6 different feature weight calculation methods. From the table, we can see that supervised term weight calculation method is not always better than unsupervised term weight calculation method, such as *tf* and $tf \times ig$ perform less well, and $tf \times idf$ which is very popular in the field of text classification is not so excellent in this experiment. However, $df \times icf \times rf$ method measures the relationship between terms and categories in the point of view whether the term shows in the positive category and negative category or not, which can maintain good classification performance in different date scale. This is because, in the traditional text classification, the training sets generally have the following characteristics: the category distribution is balanced, each document of the category can better represent this category, a document of the category is more concentrated in the arrangement of feature space. However, in practical application, the real text corpus and the use of the environment are often not satisfied with the above characteristics, therefore, the existing feature weight calculation method can't be effective in the short text.

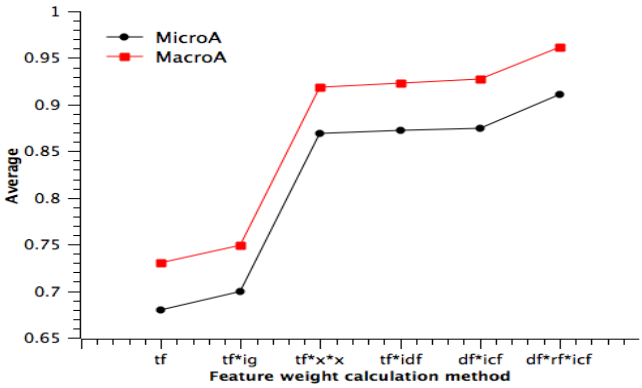

Figure 3. Comparison of the MicroA and MacroA Value

5. Conclusion

In the present study of short text classification, the feature weight calculation is still used in the traditional long text method, but the distribution of short text categories is often not so balanced, the traditional method can't get good classification results. Aiming at this problem, this paper makes an improvement on the existing feature weight calculation method, takes the correlation of terms into consideration when calculating the feature weight, and tests the performance of the method in the real corpus environment. Experimental results show that this method can improve the classification effect to a certain extent, and it is found that the classification performance of short text has not been improved greatly. The reason is the collection of the real text from the network data, including a large number of network languages and not standardized expression, resulting in the system can't accurately identify. The next step will be to introduce the method of semantic analysis to reduce the effect of the variety of words on the classification system.

ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China under grant.

References

[1] Berry MW. Survey of text mining. *Computing Reviews.* 2004; 45 (9): 548.
[2] Aggarwal C C, Zhai CX. A survey of text classification algorithms. In Mining text data. 2012; 163-222.
[3] Rao Y, Xie H, Li J, et al. Social emotion classification of short text via topic-level maximum entropy model. *Information & Management.* 2016.
[4] Sriram B, Fuhry D, Demir E, et al. *Short text classification in twitter to improve information filtering.* In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. 2010; 841-842.
[5] Wang F, Wang Z, Li Z, et al. *Concept-based short text classification and ranking.* Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM. 2014: 1069-1078.
[6] Forsyth RS, Holmes DI. Feature-finding for text classification. *Literary and Linguistic Computing*. 1996; 11 (4): 163-174.
[7] Forman G. An extensive empirical study of feature selection metrics for text classification. *The Journal of machine learning research.* 2003; 3: 1289-1305.
[8] Qi X, Davison BD. *Web page classification: Features and algorithms.* ACM Computing Surveys (CSUR). 2009; 41 (2): 12.
[9] Wang P, Xu B, Xu J, et al. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing.* 2016; 174: 806-814.
[10] Rogati M, Yang. Y. *High-performing feature selection for text classification.* In Proceedings of the eleventh international conference on Information and knowledge management. 2002; 659-661.
[11] Sriram B, Fuhry D, Demir E, et al. *Short text classification in twitter to improve information filtering.* In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. 2010; 841-842.

[12] Chen M, Jin X, Shen D. Short text classification improved by learning multi-granularity topics*. In IJCAI.* 2011; 1776-1781.
[13] Bobicev V, Sokolova. M. An Effective and Robust Method for Short Text Classification. *In AAAI.* 2008; 1444-1445.
[14] Sahami M, Heilman TD. *A web-based kernel function for measuring the similarity of short text snippets.* In Proceedings of the 15th international conference on World Wide Web. 2006; 377-386.
[15] Cavnar WB, Trenkle JM*. N-gram-based text categorization.* Ann Arbor MI. 1994; 48113 (2): 161-175.
[16] Li-guo D, Peng D, Ai-ping L. A New Naive Bayes Text Classification Algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*. 2014; 12(2): 947-952.
[17] Zhang PY. A HowNet-based semantic relatedness kernel for text classification. *Indonesian Journal of Electrical Engineering and Computer Science.* 2013; 11(4): 1909-1915.
[18] Sun A. *Short text classification using very few words.* InProceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. 2012; 1145-1146.
[19] Mihalcea R, Corley C, Strapparava C. *Corpus-based and knowledge-based measures of text semantic similarity*. In AAAI. 2006; 6: 775-780.
[20] Phan XH, Nguyen LM, Horiguchi S. *Learning to classify short and sparse text and web with hidden topics from large-scale data collections.* In Proceedings of the 17th international conference on World Wide Web. 2008; 91-100.
[21] Zhou Q, Zhao M, Hu M. Study on feature selection in Chinese text categorization. *Journal of Chinese information processing.* 2004; 18(3): 17-23.