Improvement and Application of TF-IDF Algorithm in Text Orientation Analysis

Wei Wang and Yongxin Tang*
School of Information and Electric Engineering, Hebei University of Engineering, Handan City, Hebei Province, China
*Corresponding author

***Abstract*—In this paper, on the basis of traditional TF-IDF algorithm, new and improved method is proposed. By adding the position weight coefficient and weight coefficients of word class, it can calculate the words which rely on high term frequency evenly. Experimental results showed that, the improved algorithm on the precision and recall rate are good, and it makes the selection of key collection reflect the document content orientation better.**

Keywords-selection of keyword; TF-IDF; VSM; position weight; network public opinion orientation

I. INTRODUCTION

With the advent of the web 2.0 and mobile Internet, the network is not only the important source of information, but also the platform of published their own opinion. A large number of users in microblog, forum and major e-commerce shopping platform interaction, express their own ideas. It contains a large amount of emotional orientation of information available for mining[1]. Research and develop the information is a hot issue in the field of natural language processing, text orientation analysis is one of the important link. Text orientation analysis can be widely applied in the social public opinion analysis, news commentary, event analysis, buy product, quality evaluation, stock review, evaluation of market forecast analysis, film and television, books recommended etc.[2]. The text tendency analysis research is in rapid development period, how to accurately fast analysis of the text is urgently needed to solve problems.

Network media has been considered a large media after newspapers, radio, television[3]. And because of its relative to other, the network has a virtual sex, concealment, divergent, permeability and informality, etc. Internet users can real name or anonymous on the Internet and published his own voice. This makes the network media becoming more and more important. Modern social complexity determines the various complicated social issues affecting social order. It is hard to imagine that a few years ago, now in the Internet is circulating and exposure faster and faster[4]. Strengthen the network public opinion information monitoring, grasps the public opinion in time, guide public opinion actively, is the important measures to maintain social stability and people's safety of happiness.

II. METHODS OF SELECTING THE KEYWORD

Text orientation analysis to the text, firstly, is cutting the text into the collection of words the text through the word segmentation algorithm, and preprocessing the words. Because the text after preprocessing have something big storage collection of words. If all these terms as keywords, it will produce the problems such as a very large vector dimension, storage overhead, processing speed slow, etc. But in fact, there are some terms are irrelevant with the actual category and have little effect on classification. So it needs to dimensionality reduct to the text after preprocessing. Mainly is to sort all of the terms by A feature selection method, and choose the words which have representative significance as feature words. Finally it will got the dimensionality reducted word frequency matrix and Text word frequency vector.

In the study of text feature selection, the extraction algorithm general is to construct a weighting function, evaluate characteristics of the concentration of each feature independently ,and get an assessment points. Then sort all of the points, choose features in reservation number as a subset[5]. As for how many features should be selected, and what kind of evaluation function to use, it needs to experiment according to the precision and speed of text classification for decision.

For text feature selection evaluation functions mainly include term frequency, inverted document frequency[6], IG[7], x2 statistics[8]. expected cross entropy, mutual information[9], weight of evidence text, odds ratio etc.[10] The TF-IDF algorithm is commonly used method to compute the weights of keyword, in order to describe the text characteristic more accurately.

III. THE IMPROVED TF-IDF ALGORITHM

A. Traditional TF-IDF Algorithm

The traditional TF-IDF the ideas is: if a word appears in a document with high frequency and rarely appears in other documents, then thinks that the word has the very good category to distinguish ability and suitable for classification. Commonly used TF - IDF formula is as follows:

$$w(t,d) = tf(t,d) * \log\left(\frac{N}{N_t}\right) \quad (1)$$

Among them $tf(t,d)$ represents the frequency of the keyword t appear in text d, N represents the number of full text, N_t represents the number of the texts which have the word t.

For TF-IDF formula, Its main part of the formula is same as its name, including TF(Term Frequency) and IDF(Inverse Document Frequency). Although the formula considers the word frequency and document frequency two factors at the same time, the formula itself seems a bit too simple and lacks good discrimination. As for some words that commonly used but not belong to the words for stop, they will appear in most of the articles. Therefore, the item N/N_t in the formula is not appear to be much different for them. But the frequencies of the words in the Chinese articles are generally not very high(most the TF of a word is about 10-3 orders of magnitude in a 1000-word article). This makes for a lot of words in the use of TF-IDF after come out in the formula weight value difference is very small.

In itself, IDF in essence is a kind of weighted trying to suppress the noise, and simply think text frequency small words are more important, the words with high text frequency are useless. That for most of the text information is not completely correct. The simple structure of the IDF can't make extracting keywords very effectively reflect important degree and the distribution of keywords, and make it unable to complete the weights adjustment function very well. Especially in the corpus of its kind this method has great disadvantages, and often some of the same text keywords are concealed[11].

B. *Position Weight Coefficient*

For a web document, different position words in the text to reflect effect of the document content is different. Domestic some sampling statistics, the basic coincidence rate of the title of the domestic Chinese journal of natural science papers and the text content is 98%, and the basic coincidence rate of the title of the news text and the topic is 95%. An American scholar made statistics, the sentences which can reflect the topic, 80% is at the beginning, 10% appeared in the period of the tail[12]. Therefore, this paper introduced tag weight coefficient to describe the location of the special word in the text information, shown in the table1.

Assuming that C_t refers to the keyword t in the total number of web documents, ti is the ith time word t appear in the document, K_{t_i} is the position weight coefficient of word t_i, So the key t place in the web documents all of the weighted average of the formula is as follows:

$$b_t = \frac{\sum_{i=1}^{c_t} K_{t_i}}{c_t} \tag{2}$$

TABLE I. KEY POSITION WEIGHT COEFFICIENT TABLE

Keyword tag position	weight coefficient
<TITLE>	1
<H1>	0.8
<H2>、<B>、<STRONG>	0.6
Other tag	0.5

C. *Word Class Weight Coefficient*

The selection of key are all the same part of word class evaluation function calculated weight, this method doesn't take into account the different part of speech of words to express text information gap.

In this paper, it introduces the word class weight coefficient P_t to describe speech constituent of keyword. It rules noun has a value of 2.5, verbs, adjectives, and adverbs has a value of 1, the other word classes are 0.

D. *The Improvement Word Weight Calculation Method*

When calculating the text word weight variable, introduce the balance variable in to text vector, substitute $\sqrt[n]{TF}$ for TF, to eliminate the excessive influence of term frequency and reflect the balance of weight[13].

For the keywords of a text, according to the position information of the word, the speech constituents of the word and the improved algorithm, introduce the word class weight coefficient P_t and the position weight coefficient b_t into the traditional TF-IDF formula, we can get the overall description of a new formula based on the traditional TF-IDF weighted formula, As shown in the following type:

$$w(t,d) = \sqrt[n]{tf(t,d)} * \log\left(\frac{N}{N_t}\right) * P_t * b_t \tag{3}$$

In the above formula, n(n>=1) parameter is mainly used to adjust the influence of term frequency, when the n value is small, tend to the words with high term frequency; when the n is larger value, tend to the words with low term frequency. Carried on the experiment on n=1, n=2, n=3, n=4 etc. The result is that the cubic root of TF effect is best.

Thus, we can get the weight vector of the keywords, According to the weight sort all the key words, choose the first M keywords to constitute a text vector space model to represent the document content.

IV. EXPERIMENTAL ANALYSIS

A. *Data Set*

In this paper, the experimental data is adopted COAE2015 provide Chinese text corpus for evaluation, choose positive, negative and neutral text selected each 500 as the data set, including the viewpoint about economy, entertainment, news, sports, commodity and politics. All of the documents after word segmentation, marking and handling the useless words, then test the data.

B. *Evaluation Criteria*

In this paper, the emotion tendentiousness recognition method of experimental data for evaluation is SVM algorithm, emotion tendentiousness recognition performance mainly take the following evaluation target:

$$recall = \frac{a}{a+c} \times 100\ \% \quad (4)$$

$$precision = \frac{a}{a+b} \times 100\% \quad (5)$$

$$F-measure = \frac{2 \times recall \times precision}{recall + precision} \times 100\% \quad (6)$$

Among them, a is identifying the correct number of text in the test, b is the number of the discriminant error texts, c is the number of not been discriminant out text, comprehensive evaluation F-measure is the index for describing the overall performance.

C. Experimental Results and Analysis

Preprocessing of the document by word segmentation tools ICTCLAS segmentation, and the characteristic dimension is 500. Adopt generally accepted precision, recall and F-measure value to evaluate the performance of document classification. This experiment the concrete implementation process is shown in figure 1.

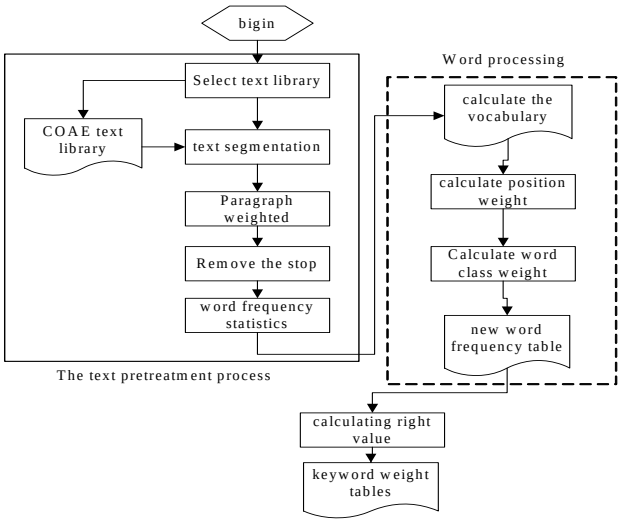

FIGURE I. FLOW CHART OF KEYWORD EXTRACTION

The experimental results are shown in table 2 and table 3.

All the data in table 2 is to calculate the average for all kinds of documents. The data in table 3 is the average precision rate and recall rate for the three algorithms based on the data in table 1. It can be seen from the experimental results, in the same conditions, using the improved formula of overall performance is better than using TF - IDF formula.

TABLE II. EXPERIMENTAL RESULTS IN VARIOUS FIELDS

Data type	Recall/%			Precision/%		
	TF	TF-IDF	Improved TF-IDF	TF	TF-IDF	Improved TF-IDF
Economy	69.11	70.01	72.23	68.53	69.52	70.33
Entertainment	69.53	69.58	71.01	69.24	70.02	70.58
News	68.99	71.32	72.33	70.12	70.35	70.68
Sports	70.32	70.65	73.56	71.01	72.05	72.82
Commodity	68.67	70.23	72.96	69.89	70.16	71.09
Politics	69.88	69.99	70.66	71.05	72.35	73.45

TABLE III. OVERALL EXPERIMENTAL RESULTS

	TF	TF-IDF	Improved TF-IDF
Recall/%	69.42	70.29	72.13
Precision/%	69.97	70.74	71.49
F-measure/%	69.69	70.51	71.81

V. SUMMARY

Aiming at the shortcomings of the traditional TF - IDF method, in this paper, it is considered the position information of keyword, word class and the balance of term frequency algorithm, to improve the TF - IDF method. Experimental results show that improved method can reflect the importance of the keyword in the text better than the traditional algorithm, and effectively improve the performance of the public opinion orientation analysis.

ACKNOWLEDGEMENT

This research was financially supported by Hebei Provincial Natural Science Foundation of China(No.F2014402093), Educational Commission of Hebei Province of China(No. ZD20131084), and the conference organizing committee.

REFERENCES

[1] Rentoumi V, Vouros G A, Karkaletsis V, et al.,Investigating Metaphorical Language in Sentiment Analysis: A Sense-to-Sentiment Perspective, Acm Transactions on Speech & Language Processing, vol.9,no.3,2012.

[2] Di Peng, Li Aiping, Duan Liguo,Based on the analysis of transition sentence text emotion tendentiousness, Computer Engineering and Design,2014.

[3] Pak A, Paroubek P., Twitter as a Corpus for Sentiment Analysis and Opinion Mining, Proceedings of the 7th Int. Conf. Language Resources and Evaluation. European Language Resources Association LREC. pp.1320-1326,2010.

[4] Chen Lidan. Public Opinion-Public Opinion Research.Beijing:China Radio and TV Press, pp.10-11,1999.

[5] Li Mingtao, Luo Junyong, Yin Meijuan, Combining with the meaning of the text of key weight calculation method, Computer Application, vol.32, no.5, pp.1355-1358,2012.

[6] Lu Yuchang, Lu Mingyu, et al., In the vector space method word weighting function analysis and structure, Computer Research and Development, pp.1205-1210, 2002.

[7] D Mladenic, M Crobelnik, Feature selection for unbalanced class distribution and Naive Bayes. In: Proc of the 16th Int'1 Conf Machine Learning(ICM199), San Francisco, Morgan Kaufmann,1999.

[8] Y Yang, J O Pedersen, A comparative study on feature selection in text categorization, In: Procof the 14th Int'1 Conf on Machine Learning (ICMU97), San Francisco, Morgan Kaufmann, 1997.

[9] Yiming Yang, Jan O.Pedersen, A Comparative Study on Feature Selection in Text Categorization. In: Douglas H. Fisherer. Proc.of The 14th International Conference on Machine Learning ICML'97, San Francisco, Morgan Kaufmann Publishers. pp.412-420,1997.

[10] Kenneth, Ward Church and Patrick Hanks,Word association norms, mutual information and lexicography. In: Proceedings of ACL27. Vancouver,Canada. pp.76-83,1989.

[11] Wang Xiaolin, Yang Lin, Wang Dong et al., Improved TF-IDF Keyword Extraction Algorithm. Computer Science.DOI:10.12677,2013.

[12] Zou Juan,Zhou Jingye, Eigenvalue extraction of synonyms processing method, Journal of Chinese Information Processing, pp.44-49,2005.

[13] Chen Keli, Zong Chengqing, Wang Xia, Based on mass balance of the real text corpus analysis and text classification method. Calculated and based on the content of text processing language--The 7th session of computational linguistics joint academic meeting,2003.