Improvement of TFIDF Algorithm based on Different Information of Text

Zhe Zhang, Zihao Su, Zhiwei Shi

College of computer technology and Engineering, Tianjin University of Technology and Education, Tianjin 300222, China.

Abstract

TFIDF algorithm is often used to extract keywords in the article, but it only considers the word frequency elements, which makes the algorithm have some defects in text classification. If a certain type is less in text, the suppression effect of IDF is not obvious. This paper presents a new improved method, which is called TFIDFZ algorithm. Then, the improved algorithm is given different weights according to the word character and different positions in the text. The improved algorithm is named TFIDFZW algorithm. The experimental results show that the precision and recall rate of TFIDFZ algorithm and TFIDFZW algorithm are better than those of traditional methods.

Keywords

TFIDF Algorithm Improvement; Text Classification; Data Mining; Precision; Recall.

1. Introduction

TFIDF algorithm is a classic feature weight algorithm, which is often used in text keyword training. Because of its relatively simple algorithm, high accuracy and recall rate, it is favored by related researchers and many application fields. In reference [1], Salton first proposed the concept of ODF (opted to document frequency), and then changed ODF to IDF (inverse to document frequency). Salton repeatedly demonstrated the effectiveness of TFIDF formula in information retrieval.

But the algorithm also exposed many shortcomings, such as the algorithm only considers the word frequency information, resulting in the selection of keywords is more limited. Many scholars put forward different improvement schemes for the problems of TFIDF algorithm. Kong qiuqiang and he Qianhua combine TFIDF with classification tree to improve efficiency [2]. Li Xueming et al. Introduced the concept of information entropy and proposed a TFIDFIGE algorithm based on information gain and information entropy [3]. Xu Fengya and Luo Zhensheng considered the distribution of feature items between and within classes, and jointly weighted the distribution information and low frequency and high weight feature information [4]. Soucy et al. Proposed a new weighting method based on the statistical estimation of word importance [5]. Xiong Zhongyang et al. Considered the lack of distribution information of feature items in inter class, intra class and incomplete classification, introduced the dispersion of feature items in inter class and intra class distribution [6]. Guo Hongyu comprehensively considered the distribution of feature items in each category during weight calculation, and proposed a weight calculation method e TFIDF based on category distribution [7]. Wang Bin, Si yangtao, Fu Juntao use TFIDF value to form feature vector as input of Bayesian algorithm to realize news text classification [8]. But Tang Peng, Xu Tiancheng, Zhang Shuhan will improve TFIDF features and combine SVM model to design an automatic Chinese text classification system [9]. Yu Wan, Gu, Yaru, Wang et al. Proposed the improvement of the dual parallel adaptive computing model [10]. Celestine iwendi, Suresh ponnan and others also applied the improved TFIDF to the text classification of large data sets [11]. Ishita daga, anchal Gupta D improved TFIDF algorithm by giving different weights according to the location information [12]. In this paper, the vector space model is used as the representation method of Web text in chong 13. TFIDF algorithm is improved. Maryam habibi [14] and others use theme modeling technology and sub module reward function to promote the diversity of keyword set, match potential theme diversity and reduce the noise of natural language recognition, so that the algorithm can extract keywords more

accurately. Li Fan et al. [15] of Tsinghua University realized a new algorithm to classify IDF function by using evaluation function instead of TFIDF algorithm. Then, the improvement of feature screening was discussed from the angle of how to relax the assumption of feature independence and use the hierarchical relation. Wang Meifang and others applied weight calculation function to feature selection, and proposed a new evaluation function based on the improved TFIDF algorithm. It introduced the category information into the feature item, extracted the feature items related to the category, and made up the defect of TF IDF algorithm [16]. It is proved that the improvement of TF IDF algorithm is a hot issue in recent years, and the improvement of TF IDF algorithm is also in a developing process.

This paper proposes a new improved method, and gives different weights according to the part of speech and the location of keywords. The improved algorithm is applied to the text classification standard data set, and good results are obtained, which verifies the effectiveness of the improved algorithm.

2. TFIDF

2.1 Algorithm Introduction

TFIDF algorithm (term frequency inverse document frequency) mainly embodies the following ideas: the higher the frequency of a word appearing in a specific document, the stronger its ability to distinguish the document content attributes (TF), the wider the range of a word appearing in a document, the lower its ability to distinguish the document content attributes (IDF).

(1) The formula (1) of TF is as follows:

$$f_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

(2) The formula (2) of IDF is as follows:

$$IDF = log\frac{|D|}{|\{j : t_i \subseteq d_j\}|} \tag{2}$$

(3) The formula (3) of TFIDF is as follows:

$$\text{TFIDF} = \text{TF} * \text{IDF} \tag{3}$$

2.2 Insufficient algorithm

TFIDF algorithm formula is relatively simple, easy to understand, and has been widely used in text similarity calculation. But the algorithm also has many shortcomings, such as: if a class has less text, the suppression effect of IDF is not obvious. In view of the above characteristics, we need to improve the algorithm and consider different text characteristics.

3. Improvement of TFIDF algorithm

3.1 TFIDFZ algorithm

TFIDF considers the document set as a whole, in which the calculation of IDF does not consider the distribution of feature items between classes and within classes. If a feature item appears in a large number in a certain category, but rarely in other categories, the classification ability of such feature item is obviously very strong, and it should be given a higher weight. However, according to the definition of IDF, if feature items appear in more documents, the IDF becomes smaller, resulting in smaller weight. According to the above thought, the improved algorithm is named TFIDFZ algorithm.

By modifying the calculation method of IDF formula (2) in TFIDF, the weight of feature items that frequently appear in a class is increased. Let D be the total number of documents, q be the number of documents containing feature t, p be the number of documents containing feature t in a class C, and r be the number of documents containing feature t except for class C. Then the IDFZ formula (4) of t in class C is as follows:

$$IDFZ = log\frac{p}{q} * |D| = \log((\frac{p}{p+r}) * |D|) \tag{4}$$

3.2 TFIDFZW algorithm

Based on TFIDFZ algorithm, the weighted processing is made according to the word character and the position information of words in text. Different weights are given to different information, named TFIDFZW algorithm.

3.2.1 Weighted by part of speech

Words in the text can be roughly divided into notional words and function words. Notional words mainly include nouns, pronouns, verbs, quantifiers and adjectives. Function words mainly include adverbs, prepositions, modal particles, exclamations and onomatopoeia. In the text processing, the most important work is word segmentation. It is not easy to do well in word segmentation. This paper uses the jieba package in Python to do word segmentation. After word segmentation, it is necessary to deal with some function words, such as modal particles, exclamations, onomatopoeia, etc.,because these words can not be used as keywords. Then according to the characteristics of general text, it may be used as keywords The nouns, verbs and adjectives of key words are given different weights in turn. The specific weight assignment corresponding table is shown, see Table 1.

Table 1. Part of speech weight correspondence table

Part of speech	Weight name	Weight
noun	n	8
verb	v	4
adjective	adj	2
Other parts of speech	Sother	1

3.2.2 Weighted by location information

According to the writing habits, most of the texts are written in the order of "total - divided - Total". That is, the first paragraph of the first paragraph usually talks about the central idea of the whole article, and the middle section is discussed and finally summarized. According to such writing habits, the corresponding words of the first and the last paragraph should be selected as the keyword words for the article. Then, the first and the last paragraphs should be given higher weights than the middle ones. This is the whole writing rule, and it is also applied to every natural paragraph, then the first and last sentences in the paragraph should also be given a higher weight. The specific weight is given, see Table 2.

Table 2. Corresponding table of position information weight

Location of occurrence	Weight name	Weight
First paragraph	FirstP	3
Last paragraph	LastP	2
First sentence	FirstS	3
Epilogue	LastS	2
Other locations	Lother	1

4. Similarity calculation

TFIDF algorithm is often combined with cosine similarity. This paper compares the similarity between texts by calculating cosine similarity. The greater the cosine similarity, the greater the similarity between texts. By adding the synonyms of keywords, more and more accurate keywords represent the text, the keyword vector constitutes the word vector of the text, and the cosine similarity is calculated to infer the similar text.

4.1 Introduction of cosine similarity algorithm

The cosine value between the word vectors of two texts can be obtained by using Euclidean dot product formula (5) The results are as follows:

$$\vec{a} * \vec{b} = ||a||||b|| \cos\theta \quad (5)$$

Given the word vectors a and B of two texts, the other string similarity θ is given by the dot product and vector length, as shown in the following formula (6):

$$similarity = cos(\theta) = \frac{A * B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} * \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (6)$$

5. Comparative analysis of experiments

5.1 Data set

In order to verify the effectiveness of TFIDFZ algorithm and TFIDFZW algorithm, experiments are conducted on standard data sets. This paper selects a common data set of text classification, which includes a total of more than 3000 articles, and initially classifies all articles into four categories: women, sports, literature and campus.

5.2 Experimental evaluation method

The experimental evaluation method compares the performance of traditional TFIDF algorithm with TFIDFZ algorithm and TFIDFZW algorithm by using the commonly used accuracy and recall evaluation indicators in the field of text classification.

The accuracy rate is for our prediction results, which means how many of the positive samples are real positive samples. Then, there are two possibilities to predict a positive class. One is to predict a true positives as a true positives (TP), and the other is to predict a false positives as a false positives (FP). The formula (7) is as follows:

$$P = \frac{TP}{TP + FP} \quad (7)$$

The recall rate is for our original sample, which indicates how many positive examples in the sample have been correctly predicted. There are also two possibilities. One is to predict the original true positives as a true positives (TP), and the other is to predict the original true positives as a false negatives (FN). The formula (8) is as follows:

$$R = \frac{TP}{TP + FN} \quad (8)$$

In addition, there is also F1 value, which is the harmonic mean of the accuracy rate and recall rate. The corresponding formula of formula (9) is:

$$\frac{2}{F1} = \frac{1}{P} + \frac{1}{R} \quad (9)$$

In the case of high accuracy and accuracy, The value will also be high. Usually, the value of F1 is used to reflect the difference of text classification performance, and its calculation formula is simple (10) For:

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (10)$$

5.3 Experimental comparison and analysis

TFIDF algorithm, TFIDFZ algorithm and TFIDFZW algorithm are used to do three comparative experiments, and the corresponding accuracy rate, recall rate and F1 value are obtained. Through the comparison of these three values, the advantages and disadvantages of their algorithms are compared. The specific experimental results are shown, see Table 3.

Table 3. Comparison results of three groups

	TFIDF algorithm			TFIDFZ algorithm			TFIDFZW algorithm		
	precision	recall	F1	precision	recall	F1	precision	recall	F1
female	0.7446	0.9459	0.8333	**0.7811**	0.9459	**0.8631**	**0.8213**	0.9459	**0.8851**
literature	0.7096	0.7333	0.7213	**0.8**	**0.8**	**0.8**	**0.8695**	**0.909**	**0.8889**
campus	0.9459	0.8139	0.875	**0.9531**	0.8139	**0.8835**	**1**	0.8139	**0.9069**
Sports	0.9107	0.8947	0.9026	0.9107	**0.9126**	**0.9111**	0.9107	**0.9226**	**0.9716**
average	0.8277	0.8469	0.8331	**0.9045**	**0.8841**	**0.8906**	**0.9429**	**0.9127**	**0.9257**

Because F1 value combines accuracy and recall, it is more objective to focus on F1 value. By comparing the F1 values of traditional TFIDF algorithm, TFIDFZ algorithm and TFIDFZW algorithm, it is concluded that TFIDFZ algorithm and TFIDFZW algorithm have better experimental results. The experimental results are shown, see Figure 1.

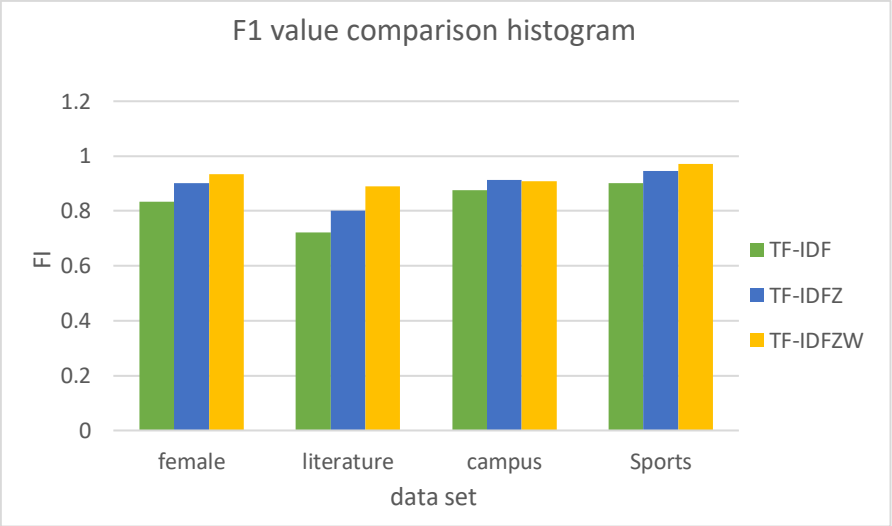

Figure 1. F1 value comparison histogram

6. Conclusion

This paper presents an improved algorithm of IDF in TFIDF. Then on this basis, we consider the part of speech, position information and add weights to improve the method. Through the experimental verification of standard data sets in text classification, the method has good results. In a large number of experiments, compared with the traditional TFIDF algorithm can be more accurate classification. Interested readers can learn from the method of this paper, for other data sets or to change, further experimental research.

References

[1] G. Salton, Clement T. Yu. On the construction of effective vocabularies for information retrieval. ACM SIGIR forum, 1973, 9(3):48-60.

[2] Kong Qiuqiang, He Qianhua. Engineering text information classification based on TFIDF and classification tree [J]. Computer applications and software, 2014, 31(06):174-176.

[3] Li Xueming, Li Hairui, Xue Liang, He Guangjun. TFIDF algorithm based on information gain and entropy [J]. Computer Engineering, 2012,38 (08): 37-40.

[4] Xu Fengya, Luo Zhensheng. Improvement of feature weight algorithm in automatic text classification [J]. Computer engineering and application, 2005 (01): 181-184.

[5] Soucy P, Mineau G.W. Beyond TFIDF weighting for text categorization in the vector space model[C]. International Joint Conference on Artificial Intelligence. Morgan Kauf-mann Publishers Inc. 2005:1130-1135.

[6] Xiong Zhongyang, Li Gang, Chen Xiaoli, et al. Improvement and application of word weight calculation method in text classification [J]. Computer engineering and application, 2008,44 (5): 187-189.

[7] Guo Hongyu. Research on feature weight algorithm based on information entropy theory [J]. Computer engineering and application, 2013,49 (10): 140-146.

[8] Wang Bin, Si Yangtao, Fu Juntao. News classification based on improved TFIDF and Bayesian algorithm [J]. Science and technology wind, 2020 (31): 9-10.

[9] Tang Peng, Xu Tiancheng, Zhang Shuhan. Chinese text classification system based on improved TFIDF features [J]. Computer and digital engineering, 2020,48 (03): 556-560.

[10] Gu Yuwan, Wang Yaru, Huan Juan, Sun Yuqiang, Xu Shoukun. An improved TFIDF algorithm based on dual parallel adaptive computing model[J]. International Journal of Embedded Systems, 2020, 13(1).

[11] Celestine Iwendi, Suresh Ponnan, Revathi M, Kathiravan Srinivasan, Chuan-Yu Chang. An Efficient and Unique TFIDF Algorithmic Model-Based Data Analysis for Handling Applications with Big Data Streaming[J]. Electronics,2019,8(11).

[12] Ishita Daga, Anchal Gupta, Raj Vardhan, Partha Mukherjee. Prediction of Likes and Retweets Using Text Information Retrieval[J]. Procedia Computer Science, 2020,168.

[13] Chu jianchong, Liu Peiyu, Wang Weiling. Improvement of word weight calculation method in Web document [J]. Computer engineering and application, 2007 (19): 192-194.

[14] Maryam Habibi, Andrei Popescu-Belis. Keyword extraction and clustering for document recommendation in conversations. 2015, 23(4):746-759.

[15] Li Fan, Lu Mingyu, Lu Yuchang. Research on new method of text feature extraction [J]. Journal of Tsinghua University (NATURAL SCIENCE EDITION), 2001 (07): 98-101.

[16] Wang Mei fang, Liu Pei Yu, Zhu Zhen Fang. Feature selection method based on TFIDF [J]. Computer engineering and design, 2007 (23): 5795-5796.