Advances in Intelligent Systems Research, volume 147
International Conference on Network, Communication, Computer Engineering (NCCE 2018)

Research on Text Classification Based on Improved TF-IDF Algorithm

Huilong Fan [1, 2, a)], Yongbin Qin [1, 2]

[1] *Guizhou Key Laboratory of Public Big Data, Guizhou University, Guiyang, 550025, P.R. China.*
[2] *College of Computer Science and Technology, Guizhou University, Guiyang, 550025, P.R. China.*

a) hlfanpro@qq.com

Abstract. In solving the problem of feature weight calculation for automatic text classification, we use the most widely used TF-IDF algorithm. Although the algorithm is widely used, there is a problem that the feature categories have different weights when calculating the weights. This paper proposes an improved TF-IDF algorithm (TF-IDCRF) that takes into account the relationships between classes to complete the classification of texts. By modifying the calculation formulas of IDF to correct the problem of insufficient classification of feature categories, the naive Bayes classification algorithm is used to complete the classification. Finally, the proposed algorithm is compared with two others improved TFIDF algorithms. The results of the three text classification evaluation indicators show that the proposed algorithm has certain advantages in text classification.

Key words: TF-IDF; text classification; Bayesian; evaluation index.

INTRODUCTION

The rapid development of information technology has promoted the explosive growth of data, and a large number of new texts have emerged. It is particularly important how to manage these huge text data. Therefore, the role of text categorization is also more and more important. The research of automatic text categorization has practical application value in the fields of data mining and text analysis. The general steps of text classification include word segmentation, feature entry selection, text representation, training model, prediction category, and so on. The selection of feature terms is the premise and basis of text classification. If the extracted feature terms cannot express the content represented by the text and the differences between different categories of documents, the final text classification result is meaningless. The number of general words in the text is many, and often contains some useless words, so it is not reasonable to use all the words in the text as classification. Generally, according to a certain screening strategy, those entries that contribute a great deal to the classification are selected to classify the text. Common term filtering strategies include TF-IDF, DF, MI, CHI, ECE, etc. [1] The TF-IDF algorithm is a common method for extracting feature entries in the text classification process, and it is simple and highly efficient.TF-IDF is a statistical method for assessing the importance of a word for one document or one of some corpuses. The importance of a word increases proportionally with the number of times it appears in the document, but at the same time it decreases inversely with the frequency with which it appears in the corpus. TF (term frequency) refers to the number of occurrences of a given word in the file. This number is usually normalized to prevent it from biasing towards long files. The improvement of TF-IDF algorithm helps to improve the accuracy of text classification results, which has practical significance in the field of data mining and artificial intelligence. At present, many scholars have improved the traditional TF-IDF algorithm or combined with other algorithms to improve the accuracy of text classification. However, most methods still have problems such as large amount of calculation and unsatisfactory classification results. The traditional TF-IDF algorithm still exists in the category. The problem of weaker distinguishing ability.

Aiming at the above problems, this paper proposes an improved algorithm TF-IDCRF, which fully considers the problem that feature entries cannot distinguish between the categories and adopts naive Bayes classifier to classify text data.

RELATED WORK

For the TF-IDF algorithm, many scholars have done a lot of improvement work. The main improvement is to improve the algorithm based on the distribution of feature words within the class and several classes. Many scholars focus on the improvement of IDF. Calculation method. G Forman [2] the use of Bi-Normal Separation (BNS) to replace the IDF part of the original TF-IDF algorithm is proposed. It is essentially based on probabilistic statistical methods to learn the significance of category distribution. Lan [3] the use of correlation frequency (RF) instead of IDF in IF-IDF is proposed to improve the recognition of text classification. Jang H [4] a new feature weighting algorithm NTFIDF is proposed, which takes into account other factors that affect the feature weights. Q Kuang [5] a new feature weighting method, IFIDFCi, is proposed in which a new weight Ci is added to represent the difference between classes based on the original TFIDF. SJ Lee, HJ Kim [6] based on the TF-IDF model, a word filtering technique called "cross-domain comparison filtering" was proposed. YS Cai, YM Huang [7] a method of webpage classification based on improved TF-IDF is proposed, and the TF-IDF weighting formula is improved by adding web tag features. JR Li, YF Mao [8] in response to the traditional TF-IDF model's inapplicability to the extraction of keywords in news advertising service modules, et al. proposed a new probabilistic model MTF-IDF to improve the accuracy of news information data retrieval. ZY Xiong, LI Gang [9] the others did not consider the issue of distribution information between categories for IDF calculations. An improved TFIDF model considering the distribution information between categories was proposed for feature selection. The KNN algorithm and genetic algorithm were used to train the classifier. KD He, ZT Zhu [10] someone proposed an improved. TF-IDF algorithm to overcome the shortcomings of the vector space model and solve the problem that the model cannot adjust the weights very well. First, the author establishes a category keyword library, and expands and repeats. Modify the weight of keywords in the document by increasing the length of the document. L Yonghe, L Yanfeng [11] in order to overcome the deficiencies of the traditional TF-IDF and its related improved algorithms, we studied how to calculate feature weights in text classification and developed a new function TW to correct feature weights. Secondly, through a comparative experiment of verifying terms CHI and TW, it is revealed that TW can increase the weight of features in a category and reduce the weight of general but not important features. W Wang, Y Tang [12] based on the traditional TF-IDF algorithm, a new improved method is proposed. By increasing the position weight coefficient of the part of speech and the weight coefficient of the word category, words that depend on the high frequency band can be uniformly calculated. X Huang, Q Wu [13] combining with the related knowledge of information theory, the distribution of keywords in the classroom was analyzed. An improved TF-IDF algorithm was proposed and applied to the calculation of the word quantity. T Xia, T Wang [14] it was found that a term with a higher frequency and close to a low dispersive distribution should have a higher weight than a less frequent and closely distributed item. Based on this assumption, Pearson's chi-square test statistic is used. Based on this, the author proposes a term weighting algorithm based on term distribution. Y Yang [15] the author improves the traditional TF-IDF method by introducing the position weights of part-of-speech weights and feature words. S Chen, Z Jin [16] using kinetic energy theorem formula, an improved TFIDF-KE feature extraction algorithm is proposed. The algorithm consists of kinetic energy and TF-IDF. Use the formula of kinetic energy theorem to evaluate the burstiness of a word and add this value to the formula. When extracting features, you can increase the weight of some important words. X Wang, J Cao [17] an improved TFIDF algorithm is proposed. The Naive Bayes classifier is used to classify texts, and iterative algorithm is used to optimize the selection of feature words. DD Xu, SB Wu [18] a new weight calculation scheme named CTF-IDF is proposed and the accuracy of the scheme is verified using a cross-validation method. L Liu, T Peng [19] a novel cluster-based approach to mobile phone reliable counterexamples (CCRNE) was proposed. In the process of establishing classifiers, a new TFIDF improved feature weighting method was proposed to reflect the importance of a word in positive and negative training examples to describe the documents in the vector space model. CH Chen [20] in the term weighting method of news articles, a distance-based term weighting method is proposed to overcome the traditional method of treating terms as noise, resulting in lower weighted defects. This approach considers a basic feature, that is, when dealing with big news that contains a lot of news, each news article must be similar or different from other articles. All news should not be deemed to contribute equally to the weighting of specific terms.

Although there are many scholars who have improved the TF-IDF method, there are still full-time fluctuations in the feature word, and there are many computational problems such as information gain, information entropy, and correlation frequency, etc., and the complexity is high [21]. To solve these problems, this paper proposes an improved TF-IDF algorithm. By considering the calculation of IDF by considering the relationship between classes, the problem of distinguishing the weights of the feature categories is resolved to improve the accuracy of text classification.

IMPROVED TF-IDF ALGORITHM

Traditional TF-IDF Algorithm

The various forms of TF-IDF weighting are often applied by search engines as a measure or rating of the degree of correlation between documents and user queries. In addition to TF-IDF, search engines on the Internet also use a link-based rating method to determine the order in which files appear in search results. In a given document, term frequency (TF) refers to the number of occurrences of a given word in the document. This number is usually normalized (the numerator is generally less than the denominator and is different from the IDF) to prevent it from biasing towards long files.

$$TF = \frac{t}{s} \tag{1}$$

Among them, t Indicates the number of occurrences of the word in the file, and s is the sum of the number of occurrences of all words in the file.

The inverse document frequency (IDF) is a measure of the general importance of a word. The IDF of a particular term can be obtained by dividing the total number of documents by the number of documents containing the term and obtaining the quotient logarithm. The high frequency of words in a particular file, and the low file frequency of the word in the entire file set, can produce a high-weight TF-IDF. Therefore, TF-IDF tends to filter out common words and retain important words.

$$IDF = \log(\frac{M}{m} + 0.01) \tag{2}$$

Among them, M represents the total number of documents in the corpus, and m represents the number of documents containing feature terms.

Insufficiency of TF-IDF Algorithm

The main idea of TFIDF is: if a word or phrase appears in an article with a high frequency TF and is rarely found in other articles, it is considered that the word or phrase has a good class distinction capability and is suitable for classification. The TFIDF is actually: TF*IDF, TF Term Frequency, IDF Inverse Document Frequency. TF indicates the frequency of entries appearing in document d. The main idea of IDF is: if the number of documents that contain the term t is less, that is, the smaller n is, the larger the IDF is, it indicates that the term t has a good category distinguishing ability. If a document C contains a number of documents for the term t is m, and the total number of documents of other classes containing t is k, obviously the number of all documents containing t is n=m+k. When m is large, n is also large. The value of the IDF obtained by the IDF formula will be small, indicating that the term t is not strong enough to discriminate. However, in fact, if an item appears frequently in a document of a class, it means that the item is a good representative of the characteristics of the text of the class. Such an item should be given a higher weight and be selected as the characteristic words of this type of text are distinguished from other types of documents, which is the deficiency of IDF.

Improvement of TF-IDF Algorithm

In most text categories, especially in multi-category text categories, for a feature term, the term may appear in multiple texts in the category, and may also appear in other categories, thus making the feature entry weights are different. The different weights will have a great impact on the classification results, and there will be fluctuations.

In response to this problem, this article has improved the IDF calculation method. A collection of categories of documents to be classified in the selected dataset $C=\{c_1 \quad c_2 \quad \cdots \quad c_m\}$, where m is the total number of categories. A class in the collection $\mathrm{c}_\mathrm{i}\ (1\le i\le m)$, the document collection in is defined as $D_i=\{d_1 \quad d_2 \quad \cdots \quad d_n\}$, where n is D_i the total number of documents.

$$IDF' = \log(\frac{\max\left(\sum_{k=1}^{D_1} t_k^1 w_k \cdots \sum_{k=1}^{D_m} t_k^m w_k\right)}{\sum_{i=1}^{m}\sum_{k=1}^{D_i} t_k^i w_k} \times n\sum_{k=1}^{a} w_k) \tag{3}$$

Among them, D_1 Representation category c_1 the collection of documents, D_m Representation category c_m the collection of documents below. t_k^1 Feature entry t in document collection D_1 B k Number of occurrences in the document, t_k^m Feature entry t in document collection D_m in the first k the number of occurrences in the document. w_k Representation k The total number of entries in the document. n Indicates feature words in the text t the total number of occurrences.

From formula (1.3) we can see that in the known data set, for feature terms t, due to $n\times\sum_{k=1}^{a} w_k$ Is a constant, so IDF' the value of is always greater than zero, and the improved calculation method eliminates the problem of poor discrimination of the original calculation method category.

EXPERIMENTAL RESULTS

Evaluation Index

For the evaluation index of the performance of text classification algorithm, there are many evaluation methods recognized by the academic community, including: Recall, Precision, and F_1 The assessed value. Recall is the ratio of the number of relevant documents retrieved and the number of related documents in the document library. It measures the recall of the retrieval system; Precision is the ratio of the number of relevant documents retrieved to the total number of retrieved documents. It is the precision rate of the retrieval system. The evaluation value is widely used in the field of information retrieval. It is the harmonic mean of the accuracy rate and recall rate and is used to measure the performance of search classification and document classification.

Data Sets

The data set uses the corpus provided by Fudan University and selects some data as experimental data. The data types include: computer, environment, agriculture, economy, politics, and sports. After word segmentation, stop word processing and other operations, the entire data set is divided into training data sets and test data sets, as shown in Table 1:

TABLE 1. Experimental Data Parameters

category	Training set	Test set
computer	693	665
surroundings	615	603
agriculture	523	499
economic	802	799
political	507	519
physical education	624	630

Experimental Results

Experiments vectorize the text and use the TFIDF algorithm, the other improved TF-IDF algorithm, and the improved algorithm TF-IDCRF proposed in this paper respectively to calculate the weights of keywords. Then use the Naive Bayes algorithm to classify the text the classification results of the three algorithms are shown in table 2, table 3, and table 4: experiments vectorize the text, using the TFIDF algorithm that introduces position weights, and the improved TF-IDF algorithm proposed in the literature. The improved TF-IDF algorithm proposed in this paper has been used to calculate the weights of keywords, and then uses naive bayes. The Si algorithm classifies the text. The classification results of the three algorithms are shown in table 2, table 3, and table 4:

TABLE 2. Text Classification Results Based on TFIDF Algorithm with Imported Location Weights

Evaluation index	TFIDF algorithm introducing position weights					
	computer	surroundings	agriculture	economic	agriculture	physical education
Recall	72.18	70.98	62.93	62.95	68.02	68.10
Precision	81.77	77.54	70.09	68.06	74.16	79.01
F-Measure	76.68	74.12	65.41	68.34	73.19	70.68

TABLE 3. Text Classification Results Based on Other Improved TF-IDF Algorithms

Evaluation index	Another improved TF-IDF algorithm					
	computer	surroundings	agriculture	economic	agriculture	physical education
Recall	73.08	74.46	69.14	68.34	71.48	75.87
Precision	84.23	83.61	78.59	73.19	81.72	85.51
F-Measure	78.26	78.77	73.56	70.68	76.63	80.40

TABLE 4. Text Classification Results Based on Improved TF-IDF Algorithm in this Paper

Evaluation index	Improved TF-IDF algorithm based on this paper					
	computer	surroundings	agriculture	economic	agriculture	physical education
Recall	75.02	75.13	70.56	72.14	73.05	78.23
Precision	86.72	85.10	79.55	75.16	82.01	87.51
F-Measure	80.12	81.01	76.25	73.09	78.31	82.07

Table 2, table 3, and table 4 are based on text classification results of the TF-IDF algorithm that introduces position weights, the latest improved TF-IDF algorithm that is proposed in other documents, and the improved TF-IDF algorithm that is proposed in this paper. From the table, we can see that the improved algorithms presented in the three algorithms show the best classification results in the six fields. It can be seen that the algorithm proposed in this paper is based on the recall rate and precision rate. F_1 The evaluation value has a satisfactory effect.

CONCLUSION

In this paper, the IDF algorithm is improved for the problem that the traditional TF-IDF algorithm lacks the capability of classifying. The corpora provided by Fudan University are used for word segmentation, stop word processing, text conversion vectors, and calculation of keyword weights. Finally, the classifier of the unknown instance is used to classify unknowns using the naive Bayesian classifier best suited for text classification. Experiments show that the idea of improvement in this paper is effective. The improved algorithm in this paper only considers how to improve the shortcomings of the improved TF-IDF algorithm and ultimately improve the classification accuracy, ignoring the calculation efficiency in the classification process. Therefore, how to improve the calculation of the algorithm while improving the final classification accuracy Efficiency is the direction for further research in the future.

ACKNOWLEDGMENTS

The research work was supported by National Natural Science Foundation of China under Grant No.61540050, Major Applied Basic Research Program of Guizhou Province under Grant No. JZ20142001, and Major Cooperation Science Research of Guizhou Province under Grant No.KH20173002.

REFERENCES

1. Kuang, Q. and X. Xu, Improvement and Application of TFIDF Method Based on Text Classification. Computer Engineering, 2006. 32(19): p. 1-4.
2. Forman, G. BNS feature scaling: an improved representation over tf-idf for svm text classification. In ACM Conference on Information and Knowledge Management. 2008.
3. Lan, M., et al. A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In Special Interest Tracks and Posters of the International Conference on World Wide Web. 2005.
4. Jiang, H. and WQ Li, Improved Algorithm Based on TFIDF in Text Classification. Advanced Materials Research, 2012. 403-408: p. 1791-1794.
5. Kuang, Q. and X. Xu. Improvement and Application of TF•IDF Method Based on Text Classification. In International Conference on Internet Technology and Applications. 2010.
6. Lee, SJ and HJ Kim, Keyword Extraction from News Corpus using Modified TF-IDF. 한국전자거래학회지제 14 권제 4 호, 2009. 14(4).
7. Cai, YS and YM Huang, Auto-Classification of Web Page Based on the Improved TF-IDF Weighting Algorithm. Journal of Mianyang Normal University, 2010.
8. Li, JR, YF Mao, and K. Yang, Improvement and Application of TF * IDF Algorithm. 2011: Springer Berlin Heidelberg. 121-127.
9. Xiong, ZY, LI Gang, and XL Chen, Improvement and application to weighting terms based on text classification. Computer Engineering & Applications, 2008. 44(5): p. 187-189.
10. He, KD, ZT Zhu, and Y. Cheng, a Research on Text Classification Method Based on Improved TF-IDF Algorithm. Journal of Guangdong University of Technology, 2016.
11. Yonghe, L. and L. Yanfeng, Improvement of Text Feature Weighting Method Based on TF-IDF Algorithm. Library & Information Service, 2013.
12. Wang, W. and Y. Tang. Improvement and Application of TF-IDF Algorithm in Text Orientation Analysis. In International Conference on Advanced Materials Science and Environmental Engineering. 2016.
13. Huang, X. and Q. Wu. Micro-blog commercial word extraction based on improved TF-IDF algorithm. In Tencon 2013 - 2013 IEEE Region 10 Conference. 2014.
14. Tian, X. and W. Tong, an Improvement to TF: Term Distribution Based Term Weight Algorithm. Journal of Software, 2011. 6(3): p. 413-420.
15. Yang, Y. Research and Realization of Internet Public Opinion Analysis Based on Improved TF - IDF Algorithm. In International Symposium on Distributed Computing and Applications to Business, Engineering and Science. 2017.
16. Chen, S. and Z. Jin, Weibo topic detection based on improved TF-IDF algorithm. Science & Technology Review, 2016. 34(2): p. 282-286.
17. Wang, X., et al. Text clustering based on the improved TFIDF by the iterative algorithm. In Electrical & Electronics Engineering. 2012.
18. Xu, DD and SB Wu, an Improved TFIDF Algorithm in Text Classification. Applied Mechanics & Materials, 2014. 651-653: p. 2258-2261.
19. Liu, L. and T. Peng, Clustering-based Method for Positive and Unlabeled Text Categorization Enhanced by Improved TFIDF. Journal of Information Science & Engineering, 2014. 30(5): p. 1463-1481.
20. Chen, CH, Improved TFIDF in big news retrieval: An empirical study. Pattern Recognition Letters, 2016. 93.
21. Wang, Q., Evaluation of Current Data Mining Algorithms. Mini-Micro Systems, 2000.