Research on Text Representation Method Based on Improved TF-IDF

FU Yu, YU You    Corresponding Author

Department of Information Security, Naval University of Engineering, Wuhan China

*Corresponding author’s e-mail:874354471@qq.com

Abstract. According to the different importance degree between feature and text, this paper proposes a method of improved traditional TF-IDF algorithm according to the position information of the words and the ability to distinguish the texts, and proposes a text representation method based on improved TF-IDF and the corresponding word vector calculation method. Finally, the effectiveness of the method is proved by experiments.

1. Introduction

In recent years, modern information technologies such as big data and cloud computing have rapidly developed. Massive data are imported into the Internet, and the model of document storage and management has also changed from traditional reliance on manual work to reliance on computers. It is a current hotspot that how to effectively manage information in the Internet and obtain effective information from it, of which text classification is the basis of text management. The basic process of text classification is shown in Fig.1.

![Figure 1 Text Classification](image)

This paper mainly improves TF-IDF method, which is one of the text representation methods.

2. Research Status

TF-IDF is a widely used text feature weight calculation method. In order to improve the accuracy of text classification, many scholars at home and abroad have improved it. Literature [1] carries out binary normal segmentation on TF-IDF through the significance relation of class distribution. Literature [2] uses correlation frequency to replace IDF in the algorithm and proposes a TF-RF weight calculation method. Literature [3] introduces the inter-class and intra-class differences of text categories into TF-IDF calculation method. Literature [4] improves TF-IDF by the theory of Chi square distribution. Literature [5-7] improves TF-IDF algorithm and proves its effectiveness through experiments.

However, the current algorithm is still not comprehensive enough in improving TF-IDF algorithm. It is easy to ignore the influence of location and other information on feature weight. According to the different importance of different feature items in text related to text categories, the traditional TF-IDF algorithm is improved according to the location information of words and the ability to distinguish text, and the calculation method of text representation method based on the improved TF-IDF is given.
3. TF-IDF algorithm improvement

3.1 The traditional TF-IDF algorithm

TF-IDF is a common feature processing method, which consists of Term Frequency (TF) and Inverse Document Frequency (IDF). The calculation method is shown in Formula (1) and Formula (2).

\[ TF-IDF(t,d) = TF(t,d) \times IDF(t) \]  

\[ IDF(t) = \log \frac{N}{N(t) + 1} \]  

Among them, \( TF(t,d) \) represents the frequency of occurrence of the word \( t \) in the text \( d \), \( N \) represents the total number of texts, \( N(t) \) represents the number of word \( t \) contained in the text, and plus 1 is to prevent the situation \( N(t) = 0 \).

However, the simple structure of TF-IDF cannot well describe the importance of feature items. TF-IDF only considers the frequency of features in the whole text, but ignores the distribution of features in the text. If a feature occurs frequently in one category but rarely in other categories, the feature should be given a higher weight. However, according to the definition of IDF, the feature may be given a lower weight, thus the weight cannot be adjusted well. And TF-IDF counts the whole text equally when it counts the features, ignoring the location information of the features. TF-IDF also affects the weight calculation results due to the difference in the number of different categories of text in the data set.

In order to describe the features in the text more fully, the factors that affect the importance of features in actual processing are mainly as follows:

- the location of the word in the text;
- the frequency of the occurrence of words in a single text;
- the number of texts in which words appear in similar texts;
- the number and frequency of words in different types of texts are different;

Considering that the different features have different contributions in different categories, this paper proposes an improved TF-IDF text representation method by combining the proposed location information and the influence of features on text. Before processing the text, it is first necessary to standardize the feature words in the text, namely, the text normalization processing \(^{[8-11]}\), so as to avoid the deviation of feature statistics caused by different forms of synonyms.

3.2 Location-based Feature Processing

In natural language processing, the content words, such as nouns and verbs, have an effect on the text semantics. However, the function words, such as prepositions and conjunctions, in the text generally have no practical effect. If such words are added to the text classification process, it will cause a lot of noise, leading to a decline in the efficiency and accuracy of text classification. Therefore, when extracting text features, such noise-prone words should be removed first.

Feature weight refers to the quantization of feature weight according to its importance, which is a key step of converting text into calculation vectors. When determining the weight of a feature, it is often necessary to comprehensively consider the information, such as the location of the word, and formulate corresponding feature weight correction rules. For example, for words that appear in important locations (titles, abstracts, etc.) but its frequency of the full text is not high, the frequency of the words should be reasonably increased so as to make the final feature weight assignment more reasonable. There are several main factors to consider:

- Title

The title is a phrase given by the author to indicate the content of the article. It is generally required to be concise and eye-catching, including many abbreviations, which are closely related to the main content of the text. The identification of the title is helpful to accurately summarize the theme of the text, and the title itself is a high generalization of the central content of text. Therefore, the correct identification of titles can improve the quality of text classification.
Abstract and the first and last sentence

According to a survey conducted by EE. Baxendale in the United States, the probability of the central sentence of a paragraph being the first sentence is about 85%, and the probability of the last sentence is about 7%. Therefore, the weights of features in special locations should be specially treated. In addition, the sentences in the first paragraph, the last paragraph, the beginning of the paragraph, the end of the paragraph, the title and the subtitle, the subtitle, etc. are often closely related to the main content of the article. Therefore, when dealing with the features of these locations, the weight should be appropriately adjusted.

Syntactic structure

In the text structure, there is a certain connection between the sentence structure and the importance of the sentence. For example, some general semantic sentences such as "in short" and "in summary" usually contain the central content of the text.

Combining the above, the specific method for determining the feature weight is as follows: Firstly, the text is segmented and normalized, and then the results of segmentation are preliminarily counted. Secondly the frequency of occurrence is corrected by using the feature location weight rule, and then the weight of each feature is calculated by using the weight determination algorithm.

3.3 The improved TF-IDF algorithm

This paper proposes an improved TF-IDF algorithm through the factors that affect the importance of features in actual processing. In a text classification task, if a word only appears in a certain type of text and is evenly distributed, it indicates that the word can directly reflect the characteristics of the text category and should be given a higher weight. If a word appears much more frequently in a certain category than in other categories and is more evenly distributed, it indicates that the word is related to the text category closely.

Assuming that the text category set is \( \{C_1, C_2, \ldots, C_n\} \), on the basis of TF-IDF, the original feature weight calculation method is improved, and the influence of the distribution of features between intra-class and inter-class is added to obtain the weight calculation method as shown in Formula (3).

\[
\delta_i = \text{softmax}(P_{i,\text{high}} \times \log \frac{\sum_{j} \text{num}_j}{\sum_{j} N_j} \times \log \frac{|C|}{|C_i|} \times (1 - \text{sigmod}(\sigma_{i,\text{max}}^2)))
\]

Among them, \( P_{i,\text{high}} \) represents the frequency in the text with the highest occurrence frequency of the word \( i \), \( \delta_i \) represents the weight of the word \( i \), \( \text{num}_j \) represents the number of word \( i \) in the document of the class \( j \), \( N_j \) represents the number of articles in the document of the class \( j \), \( |C_i| \) represents the number of text categories containing word \( i \), \( |C| \) represents the total number of text categories, \( \sigma_{i,\text{max}}^2 \) represents the distribution variance in the category with the highest occurrence frequency of word \( i \), and \( \text{softmax} \) represents the normalization function. Therefore, \( \log \frac{\sum_{j} \text{num}_j}{\sum_{j} N_j} \) represents the relationship between the text containing a certain feature and the total data set, \( \log \frac{|C|}{|C_i|} \) represents the interclass relationship of a certain feature and \( (1 - \text{sigmod}(\sigma_{i,\text{max}}^2)) \) represents the intra-class relationship of the feature.

However, there is a problem in Formula (3). when a feature appears only in a certain category of words, but its frequency of occurrence may be very low relative to the whole text, such as terminology and unknown words, the value \( P_{i,\text{high}} \) will be low, which affects the true importance of the feature. Therefore, the weight of the feature needs to be adjusted. The main adjustment objects are terminology and category-specific unknown word, namely, and words with large association with the category, the
specific methods are as follows:

Find the average probability $P_{\text{vital}}$ of words appearing only in a certain type of text. For a certain word, first judge whether it appears only in a certain type of text, that is, judge whether $\sum_{j} n_{ij}$ is equal to 1.

If it is equal to 1, adjust its weight calculation method in Formula (4).

$$\delta_i = \text{softmax}(\max(P_{r, \text{high}}, P_{r, \text{vital}}) \times \log \frac{\sum_{j} \text{num}_j}{\sum_{j} N_{ij}} \times \log \frac{|C_i|}{|C|} \times (1 - \text{sigmoid}(\sigma_{r, \text{max}})))$$  \hspace{1cm} (4)

If $\sum_{j} n_{ij} \neq 1$, the weight calculation method is shown in Formula (3).

### 4. Text Representation Based on Improved TF-IDF

#### 4.1 The word2vec text representation method

Word2vec[12] is a common word embedding model. It is a word vector training tool developed by Google in 2013. It has the characteristics of high efficiency and simplicity. The principle of Word2vec is to model the words and contexts in the text through deep learning, and finally obtain 100-300 dimensional low-dimensional vectors.

#### 4.2 Word Vector Feature Processing Based on Improved TF-IDF

In text processing, different features have different influences on the text. For example, the features related to the theme and structure of the text are more important than the general features. Therefore, it is necessary to weight the feature items of the text.

In the Word2vec word vector method, the usual method for processing features to obtain the vector representation of text is weighted average or TF-IDF method[13-14], as shown in Formula (5) and (6), respectively.

$$\text{word2vec}_\text{doc}(i) = \frac{\sum_{k=0}^{\text{word_num}(i)} \text{word2vec}(w_k)}{\text{word_num}(i)}$$  \hspace{1cm} (5)

$$\text{word2vec}_\text{TF-IDF}_\text{doc}(i) = \sum_{k=0}^{\text{word_num}(i)} \text{word2vec}(w_k) \cdot \text{TF-IDF}(w_k)$$  \hspace{1cm} (6)

Among them, $\text{word2vec}_\text{doc}(i)$ represents the vector representation of the text $i$, $\text{word_num}(i)$ represents the number of words in the text $i$, $\text{word2vec}(w_k)$ represents the number of words corresponding to words $w_k$, and $\text{TF-IDF}(w_k)$ represents the TF-IDF value of word $w_k$.

However, the weighted average method ignores that different features have different degrees of importance to the text. The TF-IDF method treats each word equally and does not consider the influence of information such as location, the inter-class and intra-class on the degree of importance of features. It only explains the ability of words to distinguish text, and cannot well reflect the degree of importance and distribution of feature words. Therefore, this paper presents a weighted word vector processing method based on improved TF-IDF, considering the word location information and the ability of distinguishing words from categories in different categories. The calculation method is shown in Formula (7)

$$\text{word2vec}_\text{level}_\text{doc}(i) = \sum_{k=0}^{\text{word_num}(i)} \text{word2vec}(w_k) \cdot \delta_{w_k}$$  \hspace{1cm} (7)
Among them, $\delta_{w}$ is same as in Formula (3) and Formula (4).

Therefore, the text classification flow based on the improved TF-IDF text representation method can be shown in Fig. 2.

Fig. 2 Text classification flow based on improved TF-IDF text representation method

5. Experiments
In order to verify the effectiveness of the feature processing method based on word location and word level, the following experiments are carried out to verify and analyze the experimental results.

Experimental environment: Anaconda3.7, Keras, Jieba, Gensim.

Experimental data: Sogou Corpus-Sogou CS, Fudan University Chinese Text Classification Corpus.

Experimental indexes: The analysis indexes of NPL processing mainly include accuracy rate, recall rate and $F1$ value. Among them, the accuracy rate is the precision rate of the test method, the recall rate is the recall rate of the test method, and the $F1$ value is the combined value of the accuracy rate and the recall rate.

Sogou Corpus-Sogou CS collects news data from 18 channels of Sohu News from June to July 2012, including URL, title, text content, etc. Data packets in the format of .dat, with a size of 1.43GB, can be downloaded from Sogou Laboratory. Select 1,000 articles from each of the four categories, finance, science & technology, sports and entertainment. Fudan University Chinese Text Categorization Corpus is a Chinese text processing corpus developed by the Natural Language Processing Team of the International Database Center of Fudan University's Department of Computer Information and Technology. Select 1,000 texts from each of the four categories of economy, politics, history and art as experimental data.

Word2vec generally has two methods to obtain eigenvectors. The one is to use the global word vector library in the existing lexicon, which is generally large in scale. For example, the Chinese Word Vector Corpus developed by the Institute of Chinese Information Processing of Beijing Normal University and DBIIR Laboratory of People's University of China \cite{15} includes training word vectors of dozens of common corpora such as Zhihu, People's Daily, Baidu Encyclopedia and Ancient Chinese. The another is to train the text corpus by training experimental data to obtain a word vector library. For the convenience of calculation, the latter method adopted in this paper obtains a local eigenvectors library by training the whole corpus.

First, the words in the text are standardized to avoid the phenomenon of inaccurate word frequency
statistics caused by different synonyms. In order to prove the reasonability and validity of this feature text representation method, a comparative experiment was conducted by weighted average word vector method (method 1) and TF-IDF weighted word vector method (method 2)\textsuperscript{16} and text representation method based on improved TF-IDF (method 3), in which 80\% of each data set was used for training and 20\% for testing. After word segmentation, the word frequency of the word segmentation result is processed by a location weighting method. In order to make the location information of words more obvious, the feature is processed by enlarging the location. According to the survey results of EE. Baxendale in the United States, and in combination with the proportion characteristics of each part of the processed text, the processing method is shown in Formula (8).

$$n_{res} = \sum_{l} n_{loc(l)} \times loc(l)$$ \hspace{1cm} (8)

Where, $n_{res}$ represents the frequency of processed words, $n_{loc(l)}$ represents the frequency of words appearing at location $l$. $loc(l)$ represents the weight value of location. The word features are pre-screened by a preset threshold method, $TL = 0.1$, $TH = 0.45$, and finally word vectors are constructed according to the text representation method based on the improved TF-IDF. Considering the relatively short news data, the word vector dimension is set to 100 and the text classifier is SVM. The experimental results are shown in Figure 3 and Table 1.

Analysis of the experimental results shows that in the two data sets used in this paper, the average accuracy, recall rate and $F1$ value of text classification based on the improved TF-IDF text representation method are higher than the weighted average sum text representation method. Therefore, the text representation method based on the improved TF-IDF can improve the quality of text feature processing and is conducive to the task of subsequent text operations.

![Fig. 3 Experimental comparison histogram of SogouCS data](image)

![Tab. 1 Comparison of Experimental Results](image)
6. Conclusion
Aiming at the unreasonable problem of traditional text feature weight calculation method, this paper proposes an improved TF-IDF algorithm according to the location information of words in text and the relationship between feature and category, and presents a weighted vector processing method based on improved TF-IDF. Experiments show that the method is more effective than the traditional feature text representation method, which lays a foundation for subsequent text categorization operations.

Acknowledgement
National Natural Science Foundation NO.61672531

References
[1] Forman G. BNS feature scaling: an improved representation over tf-idf for SVM text classification[C]//Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008: 263-270.
[2] Lan M, Tan C. L, Low H. B, et al. A comprehensive comparative study on term weighting schemes for text categorization with support vector machines[C]//Special interest tracks and posters of the 14th international conference on World Wide Web. ACM, 2005: 1032-1033.
[3] Zhang Y. F, Peng S. M, Lv J. Improvement and Application of Text-based Classification TF-IDF Method[J]. Computer Engineering,2006(19):76-78.
[4] Zhao X H. Research on feature word weighting algorithm in KNN text classification[D]. Taiyuan: Taiyuan University of Technology, 2010.
[5] He K. D, Zhu Y. T, Cheng W. Study on text classification method based on improved TF-IDF algorithm[J]. Journal of Guangdong University of Technology,2016,33(05):49-53
[6] Yang B, Han Q. W, Lei M, et al. A short text classification algorithm based on improved TF-IDF weights[J]. Journal of Chongqing University of Technology(Natural Science),2016,30(12):108-113.
[7] Lu Y. H, Li Y. F. Improved TF-IDF Algorithm for Text Feature Item Weight Calculation[J]. Library and Information Service,2013,57(03):90-95.
[8] Luo C. D, Wu X. R, Xue K, et al. A review of social text normalization research [J]. Network New Media Technology, 2017, 6 (05): 10-14.
[9] Qian T. Chinese microblog text standardization method and key technology research [D]. Wuhan: Wuhan University, 2016.
[10] Song Y J, Yu Z. H, Chen L, et al. An improved social media text normalization method [J]. Journal of Chinese Information, 2015, 29(05): 104-111.
[11] Qian T, Ji D. H, Dai W. H. Migration-based microblogging word segmentation and text normalization joint model [J]. Journal of South China University of Technology (Natural Science Edition), 2015, 43(11): 47-53.
[12] Chen L. Text representation model and feature selection algorithm [D]. Hefei: University of Science and Technology of China, 2017.
[13] Cheng L. Design and implementation of information extraction system based on improved TF-IDF algorithm [D]. Beijing: Beijing University of Posts and Telecommunications, 2019.
[14] Rahim K, Yurong Q, Sajid N. Extractive based Text Summarization Using K-Means and TF-IDF[J]. International Journal of Information Engineering and Electronic Business (IJIEEB), 2019, 11(3).
[15] Li S, Zhang Z, Hu R. F, et al. Analogical reasoning on Chinese morphological and semantic relations[C]. ACL: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2018: 138-143.
[16] Li L. Text classification based on word2vec and convolutional neural networks [D]. Chongqing: Southwest University, 2018.