Classification and analysis of literary works based on distribution weighted term frequency-inverse document frequency

Wei Dai
Shandong Management University, Jinan, Shandong Province, 250133, China

Corresponding author and e-mail: Wei Dai, 14438120110305@sdmu.edu.cn

Abstract. Term Frequency-Inverse Document Frequency (TF-IDF) is a commonly used data mining weighting technology for information retrieval. It can evaluate the importance of a word to a text and is widely used in Internet search engines. In order to improve the text analysis ability of different text types, a variety of weighting algorithms of TF-IDF have been developed. For the word analysis of literary works of various genres, this paper adopts Distribution Weighted Term Frequency-Inverse Document Frequency (TF-IDF-DW), which takes the distribution of feature items within and between classes into account, and can get better screening results. In the text classification part, via the comprehensive weight value obtained by TF-IDF-DW, and the classification results are obtained by class center vector algorithm and Bayesian algorithm. Finally, the classification performance of TF-IDF-DW algorithm is evaluated by comparison with the traditional TF-IDF method.

1. Introduction

Text classification is a classic problem in the field of natural language processing, and the related research can be traced back to the 1950s [1]. Using text analysis technology to scientifically manage and organize massive texts in the era of big data is of great significance in literature science research, social science research, science and politics research. The key technologies of text classification mainly include text preprocessing, feature dimension reduction, classification algorithm and other modules, among which the text processing module is the most important one in the research of text classification.

At present, in the field of text preprocessing, the traditional TF-IDF algorithm has high practicability. It is often used in news document classification [2], spam filtering [3], search engine recommendation system [4] and a series of text analysis with less content, single genre and obvious features.

However, for literary works with diverse genres, scarce features and different length of text content, TF-IDF algorithm is prone to the disadvantages of uneven weight of low-frequency words and insufficient weight of high-frequency words. Zhuang Zhou [5] found in his research on personalized recommendation of literary works in 2019 that the traditional TF-IDF algorithm, when the number of text quota is high, the smaller the inverse document frequency representation is, the smaller the weight of words will become, which is not suitable for the real situation. Aiming at the defects of TF-IDF's
literature text analysis, this paper selects an algorithm TF-IDF-DW which is suitable for all text analysis.

In the face of a large number and variety of literary works, TF-IDF-DW introduces the weight of location distribution to consider the distribution of feature items in each classification. Distribution Weighted (DW) can reflect the distribution of feature items within a category. If feature items appear evenly and separately in a large number of texts of this category, rather than in a single text, it can be considered that the feature item has a higher contribution to the text, and the information it contains can better reflect the text. TF-IDF-DW algorithm only provides the means to deal with word weight in text classification. It also needs the output results processed by feature selection algorithm. It is usually used as input quantity with Bayesian algorithm, class center vector method, KNN algorithm and so on. In this paper, Bayesian algorithm and class center vector method are used to analyze the word processing results of traditional TF-IDF and TF-IDF-DW algorithms to evaluate the classification effect of TF-IDF-DW algorithm.

2. Principle
TF-IDF-DW is an improved algorithm of TF-IDF algorithm. TF-IDF algorithm can be used to evaluate the importance of a single word for a literary work. The importance of this word is positively correlated with the frequency (TF) of its literary works, and negatively correlated with the frequency (IDF) of its literary works.

The implementation steps of TF-IDF are as follows:
- Calculate the frequency TF_{i,j} of each word:
  \[ TF_{i,j} = \frac{n_{i,j}}{\sum T_j} \]

  In the above formula, \( n_{i,j} \) is the number of times the word \( i \) appears in the literary work \( j \), and \( T_j \) is the number of all the words in the literary work \( j \).
- Calculate inverse document frequency IDF value:
  \[ IDF_i = \log \frac{|D|}{\sum_{t_i \in d+1}} \]

  In the above formula, \( t_i \) is the word, \( D \) is the number of literary works containing \( t_i \), and \( d \) is the literature library.
- Calculate the TF-IDF value of each word
  \[ TF - IDF_i = TF_{i,j} \times IDF_i \]

  However, the traditional TF-IDF algorithm does not consider the distribution of feature items in each literature classification, which may lead to lower weight values for some words exposed to the same classification many times, while higher weight values for some words exposed to multiple classifications. In order to avoid such errors, the traditional TF-IDF algorithm needs to be improved. The improved method is to integrate the weight value DW, which can make the distribution of words in different literature classification be considered. The improved algorithm is TF-IDF-DW algorithm.

The weight value DW indicates the position distribution of words in different literary works:
\[ DW = \max_{j=1} DD(C_j, t)CD(C_j, t) \]

The word weight value obtained by TF-IDF-DW algorithm is as follows:
\[ TF - IDF_i - DW = TF_{i,j} \times IDF_i \times DW \]

By combining the word weight value obtained by TF-IDF-DW algorithm with the class center vector algorithm or Bayesian algorithm, a literary works classifier can be formed.
2.1. Class center vector algorithm
Taking the word weight value as a multi-dimensional vector, the average value of the vector generated by the known categories of literature works in the training library is calculated and standardized. Then calculate the similarity between the vector of the literature to be tested and the vector of the training library. Let a certain type of literature in the training set be:

\[ c = \{d_1, d_2, ..., d_k \} = \{< w_{1,1}, w_{1,2}, ..., w_{m,1} >, ..., < w_{1,k}, w_{2,k}, ..., w_{m,k} > \} \]  

(6)

In the above formula, \( w_{i,j} \) denote the weight of the word \( i \) in the literary work \( d_k \), and the vector of the category \( c_i \) can be expressed as:

\[ v_{c_i} = \{< v_{w_{1,1}}, ..., v_{w_{m,1}} >, v_{w_{j,2}}, ..., v_{w_{j,m}} >, ..., v_{w_{1,m}}, v_{w_{j,m}} > \} \]

(7)

The similarity is obtained by vector angle:

\[ \cos (x, y) = \frac{\sum_{i=1}^{n} x_i \ast y_j}{\sqrt{\sum_{i=1}^{n} x_i^2} \ast \sqrt{\sum_{i=1}^{n} y_i^2}} \]  

(8)

The higher the value of similarity \( \cos (x, y) \), the higher the possibility that the literary works belong to this category.

2.2. Bayesian algorithm
For the literature to be tested, the probability of its belonging to each category is calculated, and the category with the highest probability is the literature category. The definition formula is as follows:

\[ P(Y_k \mid X) = \frac{P(XY_k)}{P(X)} = \frac{P(Y_k)P(X \mid Y_k)}{\sum_j P(Y_j)P(X \mid Y_j)} \]  

(9)

In the above formula, \( Y \) denotes literature category, \( X \) denotes words, and \( P(Y_k \mid X) \) is the probability of literature category being \( Y_k \) category under the condition of existing words \( X \).

The steps of using TF-IDF-DW algorithm to classify literary works are shown in Figure 1.
Figure 1. Flow chart of literature classification based on TF-IDF-DW algorithm.

3. Experimental design
This experiment randomly collected 8 categories of 8 000 novel texts of a novel website for classification experiment, 8 categories are fantasy, romance, martial arts, documentary, science fiction, children, myth, suspense, and each category has 1000 news samples. 900 samples were selected from each novel category as training samples, and the remaining 100 samples were used as test samples.

Using the classification performance indicators: accuracy, recall, F1 value as the evaluation criteria of the algorithm. The accuracy rate is the proportion of novels in line with the real results; the recall rate is the probability that the novels are classified correctly; the F1 value is the harmonic value of the accuracy rate and the recall rate, which is closer to the smaller of the two values.

The influence of TF-IDF-DW algorithm and different classification algorithms on the classification results using different numbers of training samples is shown in Table 1.

Table 1. Influence of different number of training samples on the results of TF-IDF-DW novel classification algorithm.

| Class center vector | Accuracy/% | recall/% | F1/% | Bayes | Accuracy/% | recall/% | F1/% |
|---------------------|------------|----------|------|-------|------------|----------|------|
| 50                  | 72.4       | 66.5     | 69.3 | 73.2  | 68.1       | 70.5     |      |
| 200                 | 82.4       | 81.3     | 81.8 | 81.9  | 81.5       | 81.6     |      |
| 500                 | 85.8       | 83.5     | 84.6 | 85.9  | 83.4       | 84.6     |      |
| 900                 | 88.1       | 85.6     | 86.8 | 88.2  | 85.4       | 86.8     |      |
Then use the same training samples as TF-IDF-DW algorithm, and repeat the classification operation with traditional TF-IDF algorithm. The influence of TF-IDF algorithm and different classification algorithms on the classification results is shown in Table 2.

**Table 2.** Influence of different number of training samples on the results of TF-IDF novel classification algorithm.

|          | Class center vector | Bayes          |
|----------|---------------------|----------------|
|          | Accuracy/% | recall/% | F1/% | Accuracy/% | recall/% | F1/% |
| 50       | 70.3       | 64.4    | 67.2 | 71.2       | 66.3     | 68.7 |
| 200      | 80.4       | 79.4    | 79.8 | 80.1       | 79.5     | 79.7 |
| 500      | 82.9       | 81.8    | 82.3 | 83.8       | 81.5     | 82.6 |
| 900      | 86.4       | 83.3    | 84.8 | 86.3       | 83.5     | 84.8 |

4. Discussion and analysis

The accuracy of novel classification method based on TF-IDF-DW combined with class center vector algorithm and Bayesian algorithm under different number of training samples is shown in Figure 2 and figure 3. The accuracy of novel classification method based on TF-IDF-DW combined with class center vector algorithm and Bayesian algorithm under different number of training samples is shown in Figure 4 and figure 5.

**Figure 2.** Accuracy of novel classification method based on TF-IDF-DW and class center vector algorithm.

**Figure 3.** Accuracy of novel classification method based on TF-IDF-DW and Bayesian algorithm.
It can be seen from the comparison between figure 2 and figure 3 and Figure 4 and figure object that with the increase of training sample number, the classification performance indicators: accuracy rate, recall rate and F1 value all show an upward trend and then tend to be stable. Taking F1 value as the basic evaluation standard, it can be seen from Figure 6 that the effect of using class center vector and Bayesian algorithm combined with TF-IDF-DW algorithm on novel classification results is similar. When the number of training samples is 900, the F1 value can reach 86.8%.

In order to compare the accuracy of TF-IDF-DW algorithm and traditional TF-IDF algorithm, the average value of F1 value of class center vector algorithm and Bayesian algorithm is taken as the evaluation standard of classification results, as shown in Table 3.
Table 3. Comparison of average F1 values of TF-IDF-DW and traditional TF-IDF under different sample numbers.

| Sample Number | TF-IDF-DW | Traditional TF-IDF |
|---------------|-----------|--------------------|
| 50            | 0.699     | 0.6795             |
| 200           | 0.7975    | 0.7975             |
| 500           | 0.846     | 0.8245             |
| 900           | 0.868     | 0.848              |
| Average       | 0.803     | 0.782              |

By comparing the average F values of TF-IDF-DW and traditional TF-IDF in Table 3, it can be seen that the average accuracy of TF-IDF-DW algorithm with DW of location distribution is 2.1% higher than that of traditional TF-IDF algorithm.

5. Conclusions
The experimental results show that when the number of training samples is highly enough, using TF-IDF-DW algorithm combined with class center vector method or Bayesian algorithm to classify novel works, the classification performance index is high, and the F1 value can reach 86.8%. Compared with the traditional TF-IDF algorithm, the TF-IDF-DW algorithm considers the weight of location distribution to have higher accuracy in the classification of literary works, which has a certain application value in the field of text classification.

References
[1] Wang Kui, Liu baisong. Review of text classification [J]. Data communication, 2019 (03): 37-47
[2] Wang Bin, Si Yang Tao, Fu Jun Tao. News classification based on improved TF-IDF and Bayesian algorithm [J]. Science and technology wind, 2020 (31): 9-10 + 17
[3] Wu Xiaqing, Wan Guojin, Li Chengwen, Lin Mengsi, Cao Shuqiang. Research on an improved TF-IDF Chinese email recognition algorithm [J]. Modern electronic technology, 2020,43 (12): 83-86
[4] Wang Guoqiao, Niu Shaozhang. Text information classification of social e-commerce based on TF-IDF [J]. Cyberspace Security, 2020,11 (12): 32-38
[5] Zhuang Zhou. Research and implementation of personalized recommendation algorithm for literary works [D]. Beijing University of Posts and telecommunications, 2019
[6] Qi Xiaojun, LAN Haixiang, Lu Hanyu, Ding Leiding, Xue Anqi. Comparative study of Bayesian, KNN and SVM algorithms in news text classification [J]. Computer knowledge and technology, 2019,15 (25): 220-222