Oceanographic Big Data Text Categorization Algorithm
Based on Improved Mutual Information

Bei LIU, Bo YIN, Zhi-qiang WEI and Yong-quan YANG*

College of Information Science & Engineering, Ocean University of China, Qingdao, China
*Corresponding author

Keywords: Oceanographic big data, Text categorization, Mutual Information, TF-IDF.

Abstract. Research on automatic text categorization of the oceanographic big data, has always been the core work of establishing an oceanographic information database based on oceanographic big data information platform. However, in face of extreme complex Internet data and the professional needs of the oceanographic field, the categorization accuracy of the traditional tf-idf algorithm is difficult to meet the demands. Based on traditional tf-idf algorithm and mutual information algorithm, I propose improved tf-idf-miow algorithm in order to meet the demands of text categorization of the oceanographic big data. Optimize the mutual information algorithm, calculate the correlation coefficient between the characteristic word and the oceanographic field. In this way, set the weight of oceanographic field: miow by this correlation coefficient, and bring miow into traditional tf-idf algorithm. The results of automatic text categorization experiments show that the recall rate of tf-idf-miow in oceanographic field is 10.33% higher than of traditional tf-idf algorithm, and the f1-score is improved by 6.92%.

Introduction

The oceanographic big data is designed to track the related website around the world, focus on online media, covering all of the oceanographic organizations, collecting a lot of news and information to establish an oceanographic information database. However, the information on the Internet is huge and complex, so automatic text categorization is the necessary means to establish the oceanographic information database. Based on mutual information algorithm, this paper introduces the weight of oceanographic field: miow (mutual information ocean weight), and bring it into traditional tf-idf algorithm, proposes the tf-idf-miow algorithm.

The Brief Introduction of Text Categorization

Text categorization[1] refers to use computer programs and related algorithms for allocating the target text to the category by comparing the contents of the target texts and training text. The training texts consisted of a given set of predefined categories and initial texts of each category. To enable the computer can correctly identify the text and calculate correlation coefficient, it is necessary to translate texts into computer-readable model. In the field of automatic text categorization and natural language processing, the vector space model (vsm) is a widely used and effective text representation model[2]. Vsm is an kind of algebraic model, first proposed In 1969. The basic idea of vsm is to segment the the text, regard the text as a collection of characteristic words that contained in the text. Text is represented by a vector so that the computer is capable of algebraic computation. Vsm represents each characteristic word as one dimension of the vector, with the weight of the word as the vector component, namely, vec(d)=(t_1,w_1; t_2,w_2;...,t_n,w_n)[3]. t_i is the No.i characteristic word of document d, and w_i is the weight of t_i. The similarity between two documents can be measured by some distance between the document vectors, which is usually using the cosine[4].
In the expression, \( d_a \) and \( d_b \) are two contrasted documents, \( m \) is the dimension of vector, \( w_{ai} \) (\( w_{bi} \)) is the No.\( i \) dimension of vector, that is, the weight of the No.\( i \) characteristic word. In the following experiment, \( vsm \) is used to represent the document, and the similarity between two documents is calculated with the cosine value.

### TF-IDF Algorithm

The weight of characteristic word is a key factor in the similarity calculation, which is directly related to the accuracy of the similarity result between texts, thus affect the text categorization effect. \( tf-idf \) has been the most widely used weight representation algorithm in \( vsm \), because of its relatively simple, high precision and recall rate. The equation of \( tf-idf \) is as follow:

\[
\begin{align*}
  w_{t,d} &= \frac{tf_{t,d} \cdot \log_2 \left( \frac{n}{n_d} + 0.01 \right)}{\sqrt{\sum_{k=1}^{m} (tf_{k,d} \cdot \log_2 \left( \frac{n}{n_d} + 0.01 \right))^2}} \\
  \text{Sim}(d_a, d_b) &= \frac{\sum_{i=1}^{m} w_{ai} \cdot w_{bi}}{\sqrt{\sum_{i=1}^{m} w_{ai}^2 \sum_{i=1}^{m} w_{bi}^2}}. 
\end{align*}
\]

In the expression, \( d_a \) and \( d_b \) are two contrasted documents, \( m \) is the dimension of vector, \( w_{ai} \) (\( w_{bi} \)) is the No.\( i \) dimension of vector, that is, the weight of the No.\( i \) characteristic word. In the following experiment, \( vsm \) is used to represent the document, and the similarity between two documents is calculated with the cosine value.

### Improvement of Weight Calculation Based on Mutual Information

#### The Brief Introduction of Mutual Information

According to the principles disclosed in the information theory, \( mutual information(mi) \) is usually used to characterize the degree of association between two random events. In the field of text categorization, \( mi \) reflects the interdependence between the characteristic word \( t \) and \( c_i \) category. \( Mi \) calculation between \( t \) and \( c_i \) is as follow:

\[
\begin{align*}
  mi_{(t,c_i)} &= \log_2 \frac{p(t,c_i)}{p(t)p(c_i)} = \log_2 \frac{p(t|c_i)}{p(t)} \\
  \text{Sim}(d_a, d_b) &= \frac{\sum_{i=1}^{m} w_{ai} \cdot w_{bi}}{\sqrt{\sum_{i=1}^{m} w_{ai}^2 \sum_{i=1}^{m} w_{bi}^2}}. 
\end{align*}
\]
mi of characteristic word t is obtained by the sum of mi between characteristic word t and each category weighted by the probability of occurrence of each category:

\[ mi_t = \sum_{i=1}^{m} p(c_i) \cdot mi_t(c_i) = \sum_{i=1}^{m} p(c_i) \cdot \log_2 \left( \frac{p(t|c_i)}{p(t)} \right) \]  \hspace{1cm} (4)

However, there are some shortcomings in the traditional mutual information algorithm. First, according to Eq.3, mi cannot reflect the difference in frequency of a characteristic word. For example, there are two characteristic words: t_1, t_2 with different frequency, when \( p(t_1|c_i)/p(t_1) = p(t_2|c_i)/p(t_2) \), \( mi(t_1,c_i) = mi(t_2,c_i) \). But a higher frequency of a characteristic word means a stronger correlation. Second, according to the Eq.4, the global mi is usually used to represent the categorization ability of a characteristic word in the entire text set, while the ability to explore the correlation between a characteristic word and the particular category is insufficient.

**Improved Oceanographic Word Weight Based on Mutual Information Algorithm**

In the field of text categorization in the oceanographic big data, whether a characteristic word is related to oceanographic field is an important factor in the categorization process. However, traditional tf-idf algorithm considers that the more concentrated characteristic word have stronger categorization ability than the scattered words, while ignoring the correlation between characteristic words and the oceanographic category. If a characteristic word meaning related to the oceanographic field, this characteristic word have a greater contribution to the categorization of oceanographic field, and it should be given a higher weight. In this paper, I propose a weight parameter miow based on the improved mutual information algorithm to explore the correlation between characteristic words and the oceanographic field, and introduce miow into traditional tf-idf algorithm to modify the weight of characteristic words. First, according to the cited literature[9], the prior probabilities and posterior probabilities of the characteristic words are introduced to reflect the dispersion and concentration of the characteristic words. Second, in order to reflect the influence of word frequency on mutual information algorithm, I introduce the frequency parameter into mutual information algorithm. The improved mi between a characteristic word t and the c_i category is expressed as:

\[ mi_{t,c_i} = \frac{\log_{10}(tf(t,c_i)+1)}{n_{c_i}} \cdot p(t|c_i) \cdot p(c_i) \cdot \log_2 \left( \frac{p(t|c_i)}{p(t)} + 1 \right) \]  \hspace{1cm} (5)

In the expression, tf(t,c_i) is the frequency of the characteristic word t in c_i category, n_{c_i} is is the number of documents in c_i category, p(c_i|t) is the probability of c_i category when t appears.

Finally, the weighting coefficients are set up to highlight the characteristics of oceanographic field. The miow weight of characteristic word t is expressed as:

\[ miow_t = 1 + \alpha \cdot p(c_i) \cdot mi_{t,c_i} - \sum_{i=2}^{m} \beta_i \cdot p(c_i) \cdot mi_{t,c_i} \]  \hspace{1cm} (6)

In the expression, c_1 is oceanographic category, c_2-c_m are categories outside the oceanographic category, m is total number of categories, \( \alpha \) and \( \beta_i \) are weight adjustment factors of oceanographic category and other categories. The tf-idf algorithm is improved into tf-idf-miow algorithm by introducing miow weight factor:

\[ w_{t,d-MIOW} = w_{t,d} \cdot miow_t \]  \hspace{1cm} (7)

The miow embodies the following ideas: the higher the mi value between characteristic word t and oceanographic category, the higher correlation between them, and a high mi value between t and other category reduces its correlation with oceanographic category. For documents that contain more oceanographic words, tf-idf-miow algorithm give these words high weight so as to increase the similarity between these documents and training documents in oceanographic category. As a result, the probability that these documents are categorized in oceanographic category is increased.
Experiment Process and Result Analysis

Experimental Data Set and Experimental Environment

The data set used in this experiment comes from the text categorization corpus provided by the natural language processing group of Fudan University and the Sogou laboratory. There are a total of 15 categories including the oceanographic category, 1720 training documents and 1746 test documents. In the experiment process, the word segmentation is processed by the open source project IKAnalyzer, and the main part is implemented in java language.

Experimental Evaluation Criteria

K-nearest neighbor (knn) is one of the most common and efficient categorization algorithm, it is widely used in text categorization. In this paper, knn is selected as the document categorization standard, and the k value is selected as 150 after experiment. At the same time, the results of categorization were evaluated by the international standard: precision rate and recall rate. For a particular category, if it is assumed that there are a documents categorized in this category, this category contains b documents, and c documents categorized success in this category, The precision rate(p) is represented as a/c and recall rate(r) is represented as b/c. Precision rate and recall rate reflect two different aspects of categorization quality, and conflict with each other. So the f1-score is usually used to evaluate the combination of them:

$$f_1 = \frac{2 \cdot p \cdot r}{p + r}$$  \hspace{1cm} (8)

Analysis of Experimental Results

Based on the selected training documents and test documents, the traditional tf-idf algorithm and the tf-idf-miow algorithm are used to calculate the weight of characteristic words respectively. Establishing the vector space model, categorizing the test documents based on the training documents. The experimental data comparisons of the two algorithms are listed in Table 1:

<table>
<thead>
<tr>
<th>Category</th>
<th>tf-idf(p/r/f1)</th>
<th>tf-idf-miow(p/r/f1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agriculture</td>
<td>0.5330/0.9496/0.6828</td>
<td>0.4653/0.9580/0.6264</td>
</tr>
<tr>
<td>Art</td>
<td>0.6667/0.7360/0.6996</td>
<td>0.7422/0.7600/0.7510</td>
</tr>
<tr>
<td>Computer</td>
<td>0.6263/0.9323/0.7492</td>
<td>0.6068/0.9398/0.7375</td>
</tr>
<tr>
<td>Economy</td>
<td>0.6143/0.6719/0.6418</td>
<td>0.5288/0.4297/0.4741</td>
</tr>
<tr>
<td>Educate</td>
<td>0.8803/0.9115/0.8957</td>
<td>0.8908/0.9381/0.9138</td>
</tr>
<tr>
<td>Environment</td>
<td>0.6164/0.8411/0.7115</td>
<td>0.6364/0.7850/0.7029</td>
</tr>
<tr>
<td>History</td>
<td>0.4844/0.6078/0.5391</td>
<td>0.4793/0.5686/0.5202</td>
</tr>
<tr>
<td>Medical</td>
<td>0.9385/0.5810/0.7176</td>
<td>0.9733/0.6952/0.8111</td>
</tr>
<tr>
<td>Military</td>
<td>0.8136/0.4800/0.6038</td>
<td>0.8167/0.4900/0.6125</td>
</tr>
<tr>
<td>Politics</td>
<td>0.6111/0.8462/0.7097</td>
<td>0.5987/0.8034/0.6861</td>
</tr>
<tr>
<td>Space</td>
<td>0.8169/0.5800/0.6784</td>
<td>0.8028/0.5700/0.6667</td>
</tr>
<tr>
<td>Sports</td>
<td>0.2222/0.0187/0.0345</td>
<td>0.5000/0.0187/0.0360</td>
</tr>
<tr>
<td>Tourism</td>
<td>0.9205/0.8100/0.8617</td>
<td>0.9268/0.7600/0.8352</td>
</tr>
<tr>
<td>Traffic</td>
<td>0.9667/0.8208/0.8878</td>
<td>0.8687/0.8113/0.8390</td>
</tr>
<tr>
<td>Ocean</td>
<td>0.9593/0.6413/0.7687</td>
<td>0.9580/0.7446/0.8379</td>
</tr>
</tbody>
</table>

The analysis of the experimental results suggests that tf-idf-miow algorithm improved the categorization quality of oceanographic category, the recall rate has risen by 10.33%. The precision rate is less affected. While the f1-score is increased by 6.92%. It is also noted that tf-idf-miow has a certain impact on the categorization of other categories, but it has a small influences, which is generally about ±3%. Only in the economic category is has a greater impact. The experimental results
show that \textit{tf-idf-miow} algorithm improves the text categorization quality in the oceanographic category compared with traditional \textit{tf-idf} algorithm.

\textbf{Summary}

Embarking from the actual demands in oceanographic category, I proposed the improved \textit{tf-idf-miow} algorithm on the basis of traditional \textit{tf-idf} algorithm. The \textit{tf-idf-miow} algorithm adjusts the weight of the characteristic word by introducing the \textit{mutual information}-related parameter, which embodies the oceanographic characteristics of characteristic words. Compared with the traditional \textit{tf-idf} algorithm, \textit{tf-idf-miow} greatly improves the categorization quality of oceanographic category. However, the \textit{miow} weights for characteristic words cause interference to the categorization of some non-oceanographic categories, resulting in a slight decrease in them. It is the focus of the author’s next work to explore the more optimized characteristic weighting algorithm, to improve the categorization quality of oceanographic while reducing reducing interference with categorization of non-oceanographic categories.

\textbf{Acknowledgment}

This work is supported by The Aoshan Innovation Project in Science and Technology of Qingdao National Laboratory for Marine Science and Technology (No.2016ASKJ07).

\textbf{Reference}


