Improved TF-IDF for We Media Article Keywords Extraction

Xinxin Guan, Yeli Li* and Hechen Gong
School of Information Engineering, Beijing Institute of Graphic Communication, Beijing 102600, China
Email: 1145512971@qq.com

Abstract. Keyword extraction is one of the work of computer text topic mining, and it is also the basis of text analysis and public opinion analysis. The keywords extracted by the traditional TF-IDF algorithm are mainly calculated based on the word frequency. The importance of other feature words with fewer occurrences and the comments of readers below the article are not considered. Aiming at the above problems, this paper improves the traditional TF-IDF algorithm, adds the part of speech and the reader's comment as the impact factor, and recalculates the weight of TF-IDF, so that the accuracy of the algorithm is improved. This paper uses the Python language programming to crawl from the media article and implement the improvement of the algorithm. Experiments show that the improved TF-IDF algorithm has significantly improved compared with the traditional TF-IDF, in terms of accuracy, recall rate, F1, MacAvg_P, MacAvg_R and MacAvg_F1.

1. Introduction
Nowadays, with the development and continuous improvement of the Internet, we can get more and more information from the Internet. At the same time, the number of self-media is gradually increasing, which leads to the continuous expansion of information in the Internet. According to IDC's estimates, the amount of resources and information on the Internet has grown exponentially, and the total amount of Internet data in the world has increased every day, with an annual growth rate of more than 50%. According to relevant experts, in 2020, the total amount of global Internet data will reach a staggering 40ZB. People who read on their mobile phones and post comments are becoming more and more accustomed to it. How to make good use of this information will serve people better and become a difficult problem to be solved. TF-IDF has been widely used as a mature algorithm in various fields of natural language processing, the most typical of which is the search engine. Although TF-IDF has been widely used, it has certain deficiencies, especially in the field of subdivision, such as keyword extraction [1].

Aiming at this problem, this paper improves the traditional TF-IDF algorithm by adding part-of-speech and reader's comments as influence factors in the algorithm, and then recalculating the weight of TF-IDF, so that the accuracy of the algorithm is improved.

2. Technical Preparations
2.1. Data Source Selection
The data source for this article is to implement news page crawling through Python language programming. According to the needs of the data in this paper, Tencent News has been crawled to get news and comments.

The author chose Tencent News (http://news.qq.com/) as the main target for crawling, because in the 2017 China Online Media Credibility Survey Report, Tencent News Network Media has been at

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
the forefront of comprehensive influence. According to the report, Tencent News has the characteristics of high user trust, strong sense of social responsibility, wide coverage and great influence [2], and the news topic has been classified in detail in the Tencent News homepage, which has alleviated the late news topic. Classification work.

Web crawlers mainly use the URL of each web page to obtain the content of the web page, and then through continuous screening, the final result is presented to the user, instead of simply obtaining the required information by manual extraction [3].

- Web page downloader [4]: Request third-party toolkit, the request library is a module for http request written in Python language, which can easily crawl web pages;
- Web Page Parser [5]: Beautiful Soup is a Python library that can extract data from HTML or XML files. BeautifulSoup4 works with Request to greatly improve crawler efficiency.

2.2. Data Preprocessing

2.2.1. Stop word Stop words refer to words that appear very frequently in the natural environment but have no material impact on the meaning of the article or page. Removing the stop word can prevent noise interference and effectively increase the keyword density.

2.2.2. Chinese word segmentation Chinese word segmentation refers to the division of a sequence of Chinese characters into individual words. Word segmentation is the process of recombining a continuous sequence of words into a sequence of words according to certain specifications. There are many tools for Chinese word segmentation, such as jieba participle, LTP, BosonNLP, IKAnalyzer, etc. The Chinese word segmentation tool used is the HanLP Chinese word segmentation by the author, which calls the HanLP Java toolkit via Python. HanLP provides a rich set of features, a stable and clear framework, and the corpus in HanLP is constantly updated to ensure the effect of word segmentation.

3. TF-IDF Algorithm

TF-IDF (term frequency-inverse document frequency) is a commonly used weighting technique for information retrieval and text mining. The main idea is that if a word is rare, but it appears multiple times in an article, then it is likely to reflect the characteristics of this article, it is more likely to reveal the topic of this article.

3.1. TF (Term Frequency)
In the pre-processing, the stop words have been filtered out, that is, the remaining words are meaningful words, then for an article, the words that appear multiple times are the more important words, namely the word frequency, the specific expression is 1, the formula is 2.

\[
TF(t) = \frac{\text{Number of time s term t appears in a document}}{\text{Total number of term s in the document}} \tag{1}
\]

\[
tf_{ij} = \frac{n_{i,j}}{\sum_{k} n_{k,j}} \tag{2}
\]

In the above formula \( n_{i,j} \), \( t_i \) is the number of occurrences of the word in the file \( d_j \), and the denominator is the sum of the occurrences of all the words in the file \( d_j \).

3.2. IDF(Inverse Document Frequency)
The word frequency in 3.1 is only limited to one article, but when a word is also a high-frequency word in an essay or database, then the word does not seem so important. To measure the importance of a word, an importance weight adjustment parameter, IDF, is introduced. The IDF size is inversely proportional to the common level of a word. The concrete expression is 3 and the formula is 4.
4. TF-IDF Improvement

In the traditional TF-IDF algorithm, only the importance of the word frequency in the article is considered, and the semantic information of the vocabulary is not considered. If this method is used only for single storage, and the improvement is not made, the keywords that may be extracted are not particularly accurate, and are not conducive to subsequent article classification. With the progress of the times, more and more articles are being published through the Internet, and there are some readers' comments on the article below. Therefore, this paper proposes an improved method for keyword extraction from self-media articles.

4.1. Algorithm Improvement Idea

This article crawled 10,000 news from Tencent News and the corresponding user comments. The core ideas of this paper mainly have three parts: 1. Calculate the ten words with the largest TF-IDF value; 2. Weight the part of each word and the number of times each word appears in the comment, and calculate each feature word. The comprehensive weight of the article; 3. According to the comprehensive weight from the high to the end of the order to determine the keywords of the media article.

- Number of times: Though we cannot overgeneralize the importance of words by the times they have been used, it does show their importance in some degree.
- Part of speech factor: According to Chinese part of speech, it can be divided into real words and virtual words. The real words include: n, v, adj, numerals, quantifiers and pronouns. Function words include: adv, prepositions, conjunctions, auxiliary words, interjections, onomatopoeia. The word-of-speech distribution is generally dominated by nouns or noun phrases, followed by verbs, numerals, adverbs and other modifiers [6]. Considering the part of speech features can effectively avoid the drawbacks of traditional linguistic methods [7].
- User Comments: There is a corresponding user comment below each article. The user's comment is also the content corresponding to the article keyword. This article mainly counts the word frequency in the comment, that is, the TF value.

4.2. Algorithm Extraction Process

According to the design of this paper, keyword extraction is divided into three steps, data crawling, data preprocessing and comprehensive weight calculation, as shown in Figure 1.
4.3. Weight Calculation

The weight calculation function proposed in this paper is as follows:

\[ F(a) = F_{tf}(a) \times F_c(a) \times F_t(a) \]  

(6)

Among them:
- \( F(a) \) is the combined weight of 10 candidate words;
- \( F_{tf}(a) \) extracts the weight of the word \( i \) for the TF-IDF;
- \( F_c(a) \) is the part of speech extracted, the noun is assigned a value of 2, the verb is assigned a value of 1.5, and the other vocabulary is assigned a value of 1.
- \( F_t(a) \) is to extract this TF value in the comment, for example \( F_t(1) = \frac{t(1)}{t(1) + \cdots + t(10)} \), \( t(1), t(2), \ldots, t(10) \) represents the number of times 10 words appear in the comments in the text.
- Finally, the three words extracted from the high-to-low order of the calculated comprehensive weights are the keywords of this paper. The part of speech and the number of occurrences in the comments are important factors in determining the keywords.

5. Experimental Result

5.1. Evaluation Standard

In this paper, we use the traditional TF-IDF algorithm and the improved TF-IDF algorithm to crawl 1000 news and corresponding user comments from Tencent News. In the 1000 news articles crawled, 100 papers were selected from the international, financial, fashion, sports, entertainment and other aspects to complete the keyword extraction, and the experimental results were evaluated.

In the comparison of results, the evaluation was performed using the precision \( P \), the recall rate \( R \), \( F_1 \), MacAvg \( P \), MacAvg \( R \), and MacAvg \( F_1 \).

The precision ratio is the ratio of the number of related documents retrieved to the total number of related and unrelated documents retrieved; The recall rate is the ratio of the number of related documents retrieved to the total number of all retrieved and unretrieved documents in the document library; The \( F_1 \) is a comprehensive indicator of the two. When the \( F_1 \) is higher, it can indicate that the test method is more effective. The calculation method is as follows:

The following is the precision \( P \) formula:

\[ P = \frac{A}{A+B} \]  

(7)

Among them:
• A represents the number of correctly x search is also x. B represents the number of correct for others, retrieved as x; The following is the recall rate R formula:

\[ P = \frac{A}{A + C} \]  

Among them:
• A represents the number of correctly x, search is also x. C represents the number of correct for x, retrieved for others;
• The following is the F1 formula:

\[ F_1 = \frac{(\alpha^2 + 1)PR}{\alpha^2P + R} \]  

In the formula, \( \alpha \) is a weighting factor that adjusts the recall and precision.

The above is just a single calculation. To see the overall calculation of the macro average, the formula is as follows:

\[ \text{MacAvg}_p = \frac{\sum_{i=1}^{N} P_i}{N} \]  

\[ \text{MacAvg}_r = \frac{\sum_{i=1}^{N} R_i}{N} \]  

\[ \text{MacAvg}_f = \frac{\sum_{i=1}^{N} F_1}{N} \]  

5.2. Comparison Of Two Algorithms

100 articles were selected from international, financial, fashion, sports, entertainment and other aspects to complete the keyword extraction, and the experimental results were evaluated. The evaluation results of the TF-IDF algorithm and the improved TF-IDF algorithm are as follows:

| Evaluation index | TF-IDF algorithm | Improved TF-IDF algorithm |
|------------------|------------------|---------------------------|
| P                | 0.46 0.48 0.52 0.53 0.57 | 0.72 0.77 0.88 0.82 0.84 |
| R                | 0.51 0.47 0.55 0.52 0.58 | 0.71 0.75 0.80 0.81 0.82 |
| F1               | 0.54 0.47 0.54 0.52 0.58 | 0.71 0.76 0.85 0.82 0.83 |
| MacAvg_P         | 0.532            | 0.806                     |
| MacAvg_R         | 0.526            | 0.778                     |
| MacAvg_F         | 0.53             | 0.794                     |

6. Summary

This paper improves the TF-IDF algorithm, introduces the frequency of words in the self-media articles and the frequency of the keywords in the comments, and improves the traditional TF-IDF
algorithm only considers the importance of the word frequency in the article, and completes the Tencent through Python. Crawling news data and validation of the improved algorithm. The experimental results are ideal. The improved algorithm has a much higher accuracy, recall rate, and F factor than the traditional algorithm. At the same time, the improved algorithm has little time difference with the traditional algorithm. Therefore, the algorithm is feasible.

7. Acknowledgments
Thanks to Beijing science and technology innovation service capability construction project (PXM2017_014223_000055), Beijing Municipal Education Committee project (KM201710015004) and Beijing Institute of Graphic Communication 2018 R&D Project (Ec201802).

8. References
[1] Liu Kan, Zhou Lihong, Chen Meng 2012 Research on Clustering of Scientific Literature Based on Key Words[J]. Library and Information Service p 6-11.
[2] Dumei Media. [EB/OL]. https://www.sohu.com/a/224314671_822760.2018.
[3] Chen Meng 2018 Design and Implementation of reptile Sina news system based on Python modern information technology p 111-112.
[4] https://www.sohu.com/a/224314671_822760.
[5] http://docs.python-requests.org/zh_CN/latest/user/quickstart.html.
[6] Qian Aibing, Jiang Wei 2008 Chinese webpage keyword extraction based on improved TF-IDF. Take the news webpage as an example. Intelligence Theory and Practice p 945-950.
[7] Huang Xuan, Li Wei 2013 Chinese keyword extraction method based on multi-features. Computer and Modernization p 15-17.