Research on an Improved Algorithm of Professional Information Retrieval System

Huajia Wang
Guangdong Polytechnic Normal University
86+18702049343.China.
910291117@qq.com

Ruo HU*
Guangdong Polytechnic Normal University
86+18928790580. China.
hu68@163.com

Hong Xu
Guangdong Polytechnic Normal University
86+18998496496.China.
xu_hjq@163.com

ABSTRACT

As the Internet develops faster and faster, resources are becoming more and more abundant. It is more and more difficult for people to retrieve the information they need from a large number of resources. The professional information retrieval system came into being. However, the current system can not retrieve resources to meet the user's requirements. In this paper, I propose a new TF-IDF information retrieval improved algorithm, which makes the resources retrieved by the information retrieval system more professional and presents people with a more accurate result. The experimental results show that the TF-IDF improved algorithm can achieve higher precision P and Recall R.

CCS Concepts

• Theory of computation →Design and analysis of algorithms →Data structures design and analysis →Sorting and searching.

Keywords

Information retrieve; TF-IDF Algorithm; Improved;Resource;

1. INTRODUCTION

With the rapid development of the Internet and the increasing resources of the Internet [1-2], accessing resources through the Internet has gradually become one of the main ways for us to obtain information [3-6]. We often search for resources through general search systems such as Google and Baidu [7]. However, such search systems have a large scope and a large amount of information, which in turn leads to insufficient depth and insufficient accuracy, and also leads to the fact that the retrieved information can not meet people’s requirements [8-9]. Therefore, the professional information retrieval systems have been born, and they are mainly for specific areas, specific groups of people and some information with specific value, which is a specific and deep retrieval system [9-12]. The information retrieved by the professional information retrieval system is professional and valuable, and can meet the needs of users. However, the resources retrieved by the current information retrieval system are also beginning to fail to meet people's needs, and many less relevant resources are also mixed into people's search results. Therefore, I propose a new TF-IDF information retrieval improved algorithm to improve the information retrieval systems, which greatly improves the accuracy and efficiency of resource retrieval.

2. TF-IDF ALGORITHM

The TF-IDF algorithm is a commonly used weighting algorithm in retrieval systems. It is actually a statistical method, which is often used to note down the importance of an article in a document set or a corpus. The importance of words increases with the number of their occurrences in the article, but it also decreases with the frequency of occurrence in the corpus or file set, which is a common algorithm for Internet information retrieval.

TF is the abbreviation of Term Frequency, which mainly refers to the times of this feature word’s appearances in the article as a percentage of the total number of words in the entire article. This value needs to be normalized to prevent the information retrieved by the retrieval system from being biased towards long words[13].The formula is shown in (1):

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (1)$$

In the above formula, the numerator $n_{i,j}$ represents the number of occurrences of the word t_i in the article d_j, and the denominator $\sum_k n_{k,j}$ represents the sum of the number of occurrences of all the words of the article d_j.

IDF is the abbreviation of Inverse Document Frequency, which means that the entropy of information contained in this feature word decreases with the frequency of occurrence in the corpus. Therefore, many people will call IDF the cross entropy of the probability distribution of keywords under a specific condition [14].The formula is shown in (2):

$$\mathrm{idf}_i = \log({}^{N}/_{n_i} + \alpha) \qquad （2）$$

In the above formula, N represents the total number of files in the corpus; n_i represents the number of all files containing this feature word; α is a constant, usually taken as 0.01.

TF-IDF is the multiplication of TF and IDF. If a word appears frequently in an article and the frequency of its occurrence in other articles is low, the word can be regarded to have a good distinction and is more suitable for classification. The formula is shown in (3):

$$\mathrm{TF-IDF} = \mathrm{tf}_{i,j} * \mathrm{idf}_i = \frac{n_{i,j}}{\sum_k n_{k,j}} * \log(\frac{N}{n_i} + \alpha) \qquad （3）$$

ICDSP 2019, February 24–26, 2019, Jeju Island, Republic of Korea.
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6204-7/19/02…$15.00
DOI: https://doi.org/10.1145/3316551.3316574

As can be seen from the above formula, the feature word is proportional to the number of occurrences of $n_{i,j}$ in an article, and inversely proportional to n_i. This reduces the influence of some common words on the weight calculation, and can successfully extract some feature words with high frequency of occurrence in this document and low frequency in other documents, which is highly representative. However, because this calculation method is given a low weight, it is a major drawback of TF-IDF.

3. IMPROVED TF-IDF ALGORITHM

The calculation method of traditional TF-IDF algorithm has low weight for some thematic feature words, which is a major disadvantage of TF-IDF. In response to this shortcoming, we need to improve it. Our way of thinking is to calculate the probability distribution of the selected feature words in each class of document C_k. For example, in a certain type of document C_k, there are m_i documents containing feature words t_i, and other types of documents have k_i documents that contains the feature word t_i. Then there are a total of $n_i = m_i + k_i$ documents containing the feature word t_i. The distribution of the feature word t_i in the class document C_k is $\frac{m_i}{m_i+k_i}$. If the $\frac{m_i}{m_i+k_i}$ is larger, the distribution of the feature words in the class document C_k is larger, which express the feature word reflects the characteristics of the class. According to the above description, the traditional IDF formula have been optimized. The optimized formula is shown in (4):

$$IDF_i = \log\left(\frac{m_i}{k_i} * \frac{N}{m_i + k_i}\right) \qquad (4)$$

N represents the total number of documents in the corpus, m_i represents the number of documents containing the feature word t_i in the class document C_k, and k_i represents the number of documents in the other class documents containing the feature word t_i.

For equation (4), assumptions can be made like this:

$$f(m_i) = \frac{m_i}{k_i * (m_i + k_i)} \qquad (5)$$

Assume $m_1 > m_2 > 0$, and was taken in the formula(5) to get the following formula:

$$f(m_1) - f(m_2) = \frac{m_1}{k_i * (m_1 - k_i)} - \frac{m_2}{k_i * (m_2 - k_i)}$$

$$= \frac{m_1 - m_2}{(m_1 - k_i) * (m_2 - k_i)} > 0 \ (6)$$

It can be seen from equation (6) that the larger the m_i, the larger the $f(m_i)$; the larger the k_i, the smaller the $f(k_i)$. Therefore, the larger the m_i value, the smaller the k_i value, which makes the value of the formula (4) larger, thus the selected feature words have strong ability to distinguish documents. In this way, the disadvantage of the feature words being given low weight in the traditional TF-IDF algorithm calculation can be avoided.

An improved formula for TF-IDF can be obtained by basing on equations (1) and (4):

$$\mathrm{TF-IDF} = tf_{i,j} * idf_i = \frac{n_{i,j}}{\sum_k n_{k,j}} * \log\left(\frac{m_i}{k_i} * \frac{N}{m_i + k_i}\right) \qquad (7)$$

4. RESULT

We often use the precision P and the recall R to show the comparison between the improved TF-IDF algorithm and the traditional TF-IDF algorithm. Precision is the ratio of the number of correctly classified documents to the number of documents that are divided into such categories. The Recall is the ratio of the number of correctly classified documents to all test documents. The formula is shown in (8), (9):

$$P = \frac{Z}{A} \qquad (8)$$

$$R = \frac{Z}{B} \qquad (9)$$

Z stands for the correct number of categ ries, A stands for the o number of documents that are divided into the class, and B stands for the number of documents for all test documents.

Our dataset is from the corpus of Sohu News Network. 1,000 news articles in six categories are obtained: economic, military, movie commentary, games, animation, and travel. 300 news items are randomly selected as test sets in each category. Others are used as a training set. The results obtained are shown in Table 1:

Table 1. Comparison of traditional TF-IDF algorithm and improved TF-IDF algorithm

test method	precision	recall
traditional TF-IDF algorithm	75.6%	74.4%
improved TF-IDF algorithm	84.9%	84.3%

This table shows the results of one experiment. The subsequent experiments are carried out to obtain more reliable experimental data, as shown in Figures 1 and 2:

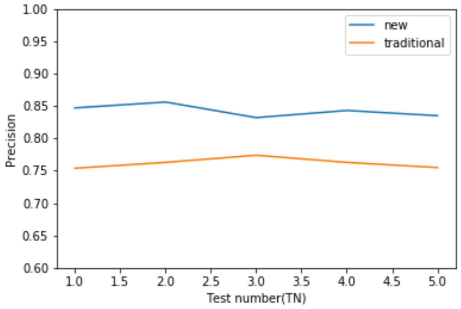

Figure 1. Comparison of precision between traditional TF-IDF algorithm and improved TF-IDF algorithm

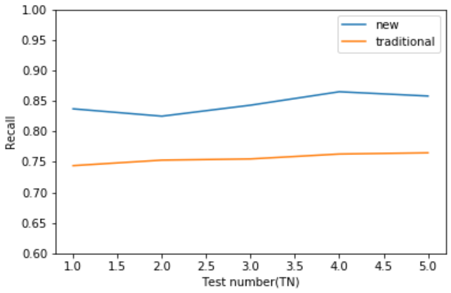

Figure 2. Comparison of recall between traditional TF-IDF algorithm and improved TF-IDF algorithm

5. CONCLUSION

Calculating the probability distribution of the selected feature words in each class document, and optimizing the IDF formula according to this probability distribution, which make the precision and recall rate of the classified documents can be increased by almost 10 percentage points. Our method can make the resources retrieved by the professional retrieval system become more precise, and greatly improves the retrieval efficiency of the retrieval system, and more popularizes the usage of the professional retrieval system.

6. ACKNOWLEDGMENTS

This study is supported by Major Project for Guangzhou Collaborative Innovation of Industry-University-Research (No.201704020196). This study is supported by Guangzhou Key Laboratory of Digital Content Processing and Security Technologies and Guangdong Provincial Application-oriented Technical Research and Development Special fund project (2016B010127006) and International Scientific and Technological Cooperation Projects of Guangdong Province (2017A050501039).

7. REFERENCES

[1] George G, Haas MR, Pentland A. Big data and management. Academy of Management Journal, 2014,57(2):321326. [doi: 10.5465/amj.2014.4002]

[2] Saint John Walker,"Big Data: A Revolution That Will Transform How We Live, Work, and Think", International Journal of Advertising-The Review of Marketing Communications,Volume 33, Issue 1:,2015,1, pp:181-183.

[3] Ruo Hu,Hui-min Zhao, Hong Xu, “A Big Data Intelligence Analysis Expression Method Based on Machine Learning”,Cluster Computing-The Journal of Networks, Software Tools and Applications, Online ISSN：1573-7543, PP:1-8, (2017-12). https://doi.org/10.1007/s10586-017-1578-9.

[4] Ruo Hu,Hui-min Zhao, Yantai Wu, “The Methods of Big Data Fusion and Semantic Collision Detection in Internet of Thing”,Cluster Computing-The Journal of Networks, Software Tools and Applications,Online:ISSN:1573-7543,pp 1-9,(2018-2). https://doi.org/10.1007/s10586-017-1577-x.

[5] Hu R, Jiang C.Y, Xu H, “A Mechanism for Healthy Big Data System Confliction Detection Using Sensor Networks”, Basic & Clinical Pharmacology & Toxicology，2016-6，vol：118，PP:42-42.

[6] Ramon Wenzel, Niels Van Quaquebeke,"The Double-Edged Sword of Big Data in Organizational and Management Research",Organizational Research Methods. Dec 2017, Vol. 30.

[7] Tim Berners-LEE, James Hendler and Oralassila,"The Semantic Web",Scientific American,Vol. 284, No. 5 (MAY 2001), pp. 34-43.

[8] Foster Provost1,Tom Fawcett, “Data Science and its Relationship to Big Data and Data-Driven Decision Making",Big Data,Volume: 1 Issue 1: 2013,2, pp:51-59.

[9] AmirGandomi, MurtazaHaider, “Beyond the hype: Big data concepts, methods, and analytics",International Journal of Information Management,Volume 35, Issue 2, April 2015, Pages 137-144.

[10] Hu R, Hu H, Xiao Z.H, “Matching Model of Health Neural Network Unit Based on Relation Object Framework”, Basic & Clinical Pharmacology & Toxicology，PP:72-73,2016-6，vol：118.

[11] Boyeong Hong, Awais Malik, Jack Lundquist, Ira Bellach, Constantine E. Kontokosta,"Applications of Machine Learning Methods to Predict Readmission and Length-of-Stay for Homeless Families: The Case of Win Shelters in New York City",Journal of Technology in Human Services.2018,1, Vol. 19: pp:1-16.

[12] Xiang L. Recommender System Practice. Beijing: Posts and Telecom Press, 2012. 40-55 (in Chinese).

[13] Hao Jiang,Wen Qiang Li. Improved Algorithm Based on TFIDF in Text Classification[J]. Advanced Materials Research,2012,1549(403).

[14] Hong Fei Sun, Wei Hou. Study on the Improvement of TFIDF Algorithm in Data Mining[J]. Advanced Materials Research,2014,3539(1042).