An improved model of document retrieval efficiency based on information theory

Li Xiaoli\textsuperscript{a}, Yin Xiaokai\textsuperscript{b}, Li Kan\textsuperscript{c}\textsuperscript{*}

Teaching and research support center of Dalian naval vessel academy, Dalian, Liaoning, China
\textsuperscript{a}email:lixiaoli136236@163.com, \textsuperscript{b}email:526037853@qq.com
\textsuperscript{c}Corresponding author:email:likandl@126.com

Abstract. Literature retrieval is an important auxiliary work of scientific research. The form of literature retrieval is becoming more and more abundant, the number of literature is huge, and the storage set is complex. The algorithm of retrieval tools directly affects the accuracy and efficiency of retrieval. Based on the analysis of the problems existing in various literature retrieval technologies, an improved TF-IDF literature retrieval efficiency model is proposed. It provides a new idea for the design and implementation of literature retrieval system.

1. Introduction
The related concepts of literature retrieval are described.

1.1. essence of literature retrieval
Literature refers to the important books and materials related to a certain subject with communication value, and the volume of literature is huge. Retrieval refers to the process of searching and obtaining specific related documents from a large number of organized related documents according to certain methods. Information retrieval is the process of obtaining documents according to the needs of study and work[1]. With the development of modern network technology, literature retrieval is more completed by computer technology. The modern literature retrieval is a process of literature retrieval by means of computer technology, using specific information retrieval model and data mining technology to obtain the target literature from the electronic information system. The efficiency of literature retrieval is closely related to the thinking logic of retrieval tools. In the era of knowledge explosion[2], the efficiency of literature retrieval is related to the ability of literature dissemination and scientific progress. It is important to study and improve the literature retrieval model[3].

1.2. corpus and measurement of relevance
Corpus refers to a large-scale electronic text database which has been scientifically sampled and processed. With the help of computer analysis tools, researchers can carry out relevant language theory and application research[4]. Corpus is the basic resource of language knowledge, but it is not equal to language knowledge. Real corpus needs to be processed (analyzed and processed) to become a useful resource. TF-IDF (term frequency – inverse document frequency) is a common weighting technique for information retrieval and data mining. TF is term frequency and IDF is inverse document frequency. TF-IDF is a statistical method to evaluate the importance of a word to a file set or one of the files in a corpus. The importance of a word increases with the frequency of its appearance in the document, but...
decreases with the frequency of its appearance in the corpus. Various forms of TF-IDF weighting are often used by search engines as a measure or rating of the relevance between files and user queries[5].

1.3. Literature search and TF-IDF
The concept of TF-IDF was first proposed by Spock Jones of Cambridge University. In 1972, Spock Jones put forward the concept of IDF in a paper entitled "special statistical interpretation of keywords and his application in literature search". Later, scholars of information theory found and pointed out that the concept of IDF is the cross entropy of the probability distribution of keywords under a specific condition. In this way, the measurement of the relevance of information retrieval is back to the information theory.

2. Principles of literature retrieval system

2.1. The realization conditions of literature retrieval
Document retrieval system realizes the function of the system through the two reciprocal processes of document storage and document retrieval. In a word, the process of literature retrieval is the comparison between the query mark and the document mark in the retrieval tool library. Effective literature retrieval is the unification of the two marks. Storage is to gather a large number of disordered literature, according to the appearance and content characteristics of the literature source, through sorting, classification[6], concentration, indexing and other processing, to make it systematic and orderly, and according to certain technical requirements to build a database or retrieval system with retrieval function, for people to search and use. The retrieval refers to the use of a good search tool or search system to find out the specific information to meet the user's requirements.

The first paragraph after a heading is not indented (Bodytext style).

2.2. Theoretical model of literature retrieval
The lack of semantic information will lead to the wrong description of scientific and technological literature retrieval data. The data structure of different sources will make the difference and repeatability of data information. This is due to the semantic deviation caused by semantic association information, resulting in the duplication of retrieval records. In order to improve the retrieval quality, this paper analyzes the traditional scientific and technological literature retrieval technology. This paper uses semantic based scientific literature retrieval mechanism to remove the shortcomings of traditional retrieval model, combines retrieval content with semantic association algorithm, introduces semantic similarity of concepts and semantic reasoning based on principles, infers appropriate object

![Figure 1 Principle block diagram of literature retrieval](Image)

Figure 1 Principle block diagram of literature retrieval
attributes for ontology construction and rule making, and constructs retrieval prototype. Through the
test results of the system, the validity and feasibility of the model are verified, which has a positive
significance for the practical research[7].

The principle of literature retrieval system is shown in Figure 1, and the retrieval model is the dotted
line in the figure.

3. Literature retrieval technology and problem description

Information retrieval is a widely used technology at present. Paper retrieval and search engine belong
to the category of information retrieval. Usually, people abstract the problem of information retrieval as:
on the document set D, for the keywords w [1] The query string Q composed of w [k] returns a related
document list d 'sorted by the matching degree (Q, d) between query Q and document D. For this basic
problem, there are various classical information retrieval models, such as Boolean model, vector model
and so on. They put forward their own solutions from different angles. Boolean model is based on
Boolean operation of set and has high query efficiency, but the model is too simple to sort different
documents effectively and the query effect is poor. In vector model, documents and query strings are
regarded as multidimensional vectors composed of words, and the correlation between documents and
queries is corresponding to the angle between vectors. However, due to the large number of words, the
vector dimension is very high, and a large number of dimensions are 0, the effect of calculating vector
angle is not good. In addition, the huge amount of computation also makes the vector model almost
impossible to implement on the massive data set such as Internet search engine. At present, TF IDF
model is widely used in search engines and other practical applications. The main idea of TF IDF model
is: if the word w appears frequently in one document D and rarely in other documents, it is considered
that the word w has a good ability to distinguish article d from other articles. The disadvantages are: (1)
only "word frequency" is used to measure the importance of words, and the subsequent document
eigenvalue sequence is composed of independent words, which can not reflect the sequence information;
(2) it is easy to be affected by the skewness of the data set, such as too many documents in a certain
category, which will lead to the underestimation of IDF; (3) the processing method is shown in the
following formula to increase the category weight; (3) the distribution deviation within and between
categories (used for feature classification) is not considered When selecting. Improvement: add the
weight of the number of documents whose current category contains term1 in IDF.

4. Improve the model process

In the era of high-speed information, the traditional retrieval technology is becoming more and more
inefficient to retrieve the content in line with their own search from a large number of massive
information data. The traditional information retrieval method is generally based on keyword matching,
and the results obtained by this matching method are mostly loose and unsystematic. This means of
acquiring knowledge is often mechanical and non intelligent. In order to solve the above problems of
literature retrieval, semantic retrieval is used to obtain accurate literature retrieval results.

The weight of each key word w in a query should reflect how much information the word provides
for the query. A simple way is to use the amount of information of each word as its weight, that is

\[ I(w) = \log P(w) \log \frac{N}{TF(w)} = -\frac{TF(w)}{N} \log \frac{TF(w)}{N} = \frac{TF(w)}{N} \log \frac{N}{TF(w)} \]  

(1)

\( N \) is the size of the whole corpus, which can be omitted. The above formula can be simplified to

\[ I(w) = TF(w) \log \frac{N}{TF(w)} \]  

(2)

However, the formula (2) has a defect: the frequency of two words is the same, one is a common
word in a specific article, and the other is scattered in multiple articles, so obviously the first word has
higher resolution.

If you make some ideal assumptions,

1) The size of each document is basically the same, which is M words, namely

\[ M = \frac{N}{D} = \frac{\sum_{w} TF(w)}{D} \]
2) Once a keyword appears in a document, no matter how many times it appears, its contribution is equal. Such a word either appears \( c(w) = \frac{TF(w)}{D(w)} \) times or zero in a document. Note that starting from \( c(w) < M \), (2), we can get the following formula:

\[
TF(w) \log \frac{N}{TF(w)} = TF(w) \log \frac{MD}{c(w)D(w)} = TF(w) \log \left( \frac{D}{c(w)D} \frac{M}{c(w)} \right) \tag{3}
\]

In this way, we can see that the difference between TF-IDF and the amount of information is the second in formula (3). Because \( c(w) < M \), the second term is greater than zero, which is a decreasing function of \( c(w) \). Rewrite the above formula as

\[
TF-IDF(w) = I(w) - TF(w) \log \frac{M}{c(w)} \tag{4}
\]

It can be seen that the more the amount of information \( I(w) \) of a word, the greater the TF-IDF value; at the same time, the more the average times of \( w \) in \( w \) hit literature, the smaller the second term, the greater the TF-IDF value. These conclusions are completely consistent with information theory.

5. Conclusions

TF-IDF is a measure of the importance of search keywords, and has a strong theoretical basis. TF-IDF model is an information retrieval model widely used in search engine and other practical applications, but there are various doubts about TF-IDF model. In this paper, a box ball model based on conditional probability is proposed for information retrieval. The core idea is to transform the problem of "matching degree between query string \( Q \) and document \( D \)" into the problem of "conditional probability that query string \( Q \) comes from document \( D \)". From the perspective of probability, it defines a clearer goal for information retrieval than the matching degree expressed by TF-IDF model. Based on TF-IDF, this paper makes some improvements and fine-tuning. It is not far from TF-IDF in principle to directly use the improved TF-IDF search engine to measure the importance of keywords.

Compared with the mechanical retrieval method, according to the semantic retrieval, it has intelligent reasoning retrieval, reasonably expands the concept of the retrieval results, and greatly improves the recall and precision of the retrieval results. At the same time, with the rapid development of various crawler technologies, the search engine of full-text retrieval is also widely used in the retrieval process, which is based on ontology technology. The research results on the idea of secondary semantic similarity are also very fruitful.

References

[1] CAO, Y., JIA Y. C., WANG, Z. (2019) Research on semantic-based scientific and technological literature retrieval technology. Microcomputer Application, 35:16-18.

[2] WANG, Y., (2020) Design of the balance allocation algorithm of document resource based on TRS information retrieval. Computer Simulation, 37:445-449.

[3] MA, Z.Q., LIAO X. h., DENG E., XIAO, W. c.,(2020) Research on comparison algorithm of similar contents. Computer and Digital Engineering, 48:2843-2849.

[4] KE, P., GAO, J.(2018) Introduction to Information Management. Science Press, Beijing.

[5] WU, J., (2017) Beauty of Mathematics. Posts & Telecom Press, Beijing.

[6] XU, F., WANG, S. C., YANG W. X., (2019) Cloud resource scheduling algorithm based on Game Theory. Computer Science, 46:295-299.

[7] QIAN, Z. H., HU, L. S., TIAN, C. S., (2019) Research on D2D multi-multiplex communication resource blocks allocation algorithm based on unbalanced solution. Journal of Electronics & Information Technology, 41:2810-2816.