Analysis and Optimization of Information Retrieval Algorithms for Unstructured Data

Kunying Li, Dexin Qiao, Xiaolian Li, Yu Ding

Computer Application Technology Department, PetroChina Research Institute of Petroleum Exploration & Development, Beijing 100083, China

Abstract. The Internet has diversified in the form of an explosion in recent years. It has spawned countless forms of Internet branching, and at the same time brought information to the PB level, and massive data is also called big data. More than 85% of the collected data is composed by unstructured and semi-structured data; in order to solve the data group management in the contract system of a large-scale energy enterprise, it aims to realize the interconnection of upstream business data, technology interoperability, research collaboration, and promote the demand for intelligent and massive unstructured data retrieval. This paper proposes a non-institutional data retrieval optimization algorithm based on periodic data heat and category labels. The algorithm is implemented by correlating the user's retrieval behavior in the cycle and combining the defined file category tags. The experimental results show that the method not only can effectively filter and sort unstructured data, but also it can provide strong support for subsequent big data analysis and edge calculation.

Keywords: Information Retrieval; unstructured data; user behavior; file category; optimize algorithm.

1. Introduction

Under the booming of the Internet, data and information have gradually become a valuable resource wealth. At present, helping traditional enterprises to realize informatization in a big data environment is the most important direction. How to obtain truly valuable information in massive data has become a particularly important topic in various industries. In order to enhance the retrieval effect of unstructured data, there are many targeted search optimization methods. Some of them use the content input by the user to judge or expand the semantics on the basis of this, in order to achieve the purpose of expanding the search scope; Some of them make special processing and distinguishing by weighting the key information of the data; others make the system record the function of the user's search history and learn to improve the retrieval efficiency. There are also methods of making the system to record history of user's search and learn to improve retrieval efficiency. In fact, if the system can grasp the user's search scope, direction and correlate its recent search content, and advancing rank of the data that the user wants or cares about, Obviously, it can improve the accuracy and experience of the user query in data filtering. Based on the calculation of term frequency and inverse document frequency, this method records and analyzes the search behavior of users in the period, and then combines the similarity between file categories to propose the non-structure based on periodic data heat calculation and associated category labels. And compare this algorithm with the existing search method. The experimental results show that the optimized algorithm is closer to the user's true search intent.

2. Research Background

The current screening analysis of unstructured data is mainly reflected in the following aspects.[1] it proposes the ontology learning method based on statistics, the relevance the domain is defined by comparing the domain corpus and compositive corpus.[2]Combine ant colony algorithms to enable system more quickly and accurately to find the data in the cloud,[3] put forward a method of calculating the weight of keywords sequence which help system to find relevant similarity of data,[4] Full-text search service provided by existing relational database is difficult to retrieve binary text, then using open source database (Hibernate) to build a retrieval model to solve this problem,[5] It is proposed to use the computer to realize the Chinese resource sharing function in the language library,
and use the T-L conversion algorithm to quickly find the target,[6] using a type of double-sorted cross-correlation successor tree binary plus verification search and pre-processing interval table search algorithm are used to improve retrieval efficiency. The above researches are involved in predictive search, content similarity and fast positioning, but do not give a specific algorithm on date Filtering. Based on the existing research, this paper further combines the periodic heat between data and assigns category attribute labels to the data, in order to improve the user's final search result ordering.

In addition, in the process of unstructured data analysis, the establishment of data model is a tool and method to realize abstract description to real things. It is a kind of mapping that represents the relationship between things in the real world through abstract forms, so a good data model directly affects the efficiency of data processing in the future. In the traditional data model design, commonly used a type of relational database to establish the relationship between primary and foreign keys on tables, every table structure is very close to physical logic. However, the cost for the company to maintain data consistency is huge which performance of read and write is relatively poor, and the cost of paying large amounts of data has also multiplied. Therefore, this paper uses a new data model based on inverted index rules, it provides a key modeling foundation for building unstructured data filtering analysis algorithms.

3. Content Relevance Ranking Algorithm and Optimization Method

Due to the increasing scale of users and projects, otherwise different users have different search angles, sensitivity and granularity of requirements. As a result, the filtering algorithm needs to provide different satisfaction to the data from different sources. We need to know what kind of data is the most important for users. The current solution is to perform preliminary processing on the data. The calculation of index process is a key value, and it is also the most common and efficient way. The processed data can effectively improve the accuracy and efficiency of the query. However, in the actual retrieval process, the user's attention and the search scope of the document are difficult to express in a single vocabulary. Not only the simple algorithm pre-processing calculation does not solve the problem of increasing data range from the root cause, but also the user community’s popularity of certain documents is also a feature of the current search experience.

3.1 Basic Correlation Evaluation Method of Data Analysis

Therefore, weight analysis design is the core premise of the algorithm after data collection. [7] The following are the two basic frequency calculation methods used in the project.

(1) Term frequency (TF): Indicates how often a given term appears in the document

TF indicates how often a given term appears in the document. With the query term \( i \) and the document \( j \), the frequency of the query term \( i \) in the document \( j \) can be defined as following mathematical expression:

\[
\text{tf}_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}}
\]

(1)

The numerator \( n_{i,j} \) represents the number of occurrences of the term \( i \) in the document \( j \), and the denominator represents the sum of the number of occurrences \( n_{k,j} \) of all the terms \( k \) in the document \( j \).

(2) Inverse document frequency (IDF): It refers to the frequency of documents in which a word appears in all documents. A larger value indicates that the vocabulary has a good class distinguishing ability and can better represent the characteristics of this type of document.

Setting the total number of all documents in the database as \( n_d \), the denominator is the number of files containing the term \( t_i \) in all documents \( d \), and then the logarithm of the quotient is expressed as the IDF frequency. However, in practice, some invalid words or unfamiliar words are often considered as keywords, which affect the overall scoring standard. Therefore, the algorithm corrects
the problem and sets invalid terms in advance, then the remaining standard vocabulary is $t_{i-m}$, and finally according to the rules of the log function:

$$idf_i = \log\left(\frac{n_d}{[j: t_{i-j} \in d_j]}\right) \tag{2}$$

Based on the above two calculation frequencies, it can be used to evaluate the importance of vocabulary for one of the files in the file library, also known as TF-IDF. The importance of vocabulary rises in proportion to the number of times it appears in the content. At the same time, the frequency of vocabulary which exist in the database declines in opposite proportions.

$$tf - idf_{ij} = tf_{i,j} \times idf_i = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log\left(\frac{n_d}{[j: t_{i-j} \in d_j]}\right) \tag{3}$$

According to the high weight tf-idf generated by the above method, not only the vocabulary without actual semantics is filtered out after optimization, also it improves the screening opportunity for important words.

### 3.2 Information Retrieval Optimization Algorithm based on Heat Computing and Category Labels in the Period of Time

[8] In the study of human behavior, it is shown that the human behavior process can be treated as a series of transactions, and it is preferred to concentrate on these tasks within a certain period of time, therefore, the user's search behavior over a period of time can also be expressed as a direct or indirect association with recent tasks. Based on the initial score of tf-idf algorithm, this paper proposes a joint optimization improvement evaluation algorithm based on data association heat and category label. Calculate the similarity to the current task in the phase time by the result attribute of the user retrieval target, and the additional score value of the document is calculated according to the correlation between data access frequency, access duration and target data for final feedback. Therefore, the scores of the data are closer to the user-oriented evaluation criteria on the system, and the search accuracy is further optimized and improved.

The basic idea of the optimization algorithm is to first locate the category keywords of the data, extract the category keywords from the data, such as: finance, tax, agriculture, information, etc., form the document category label, and obtain the user's recent query task target result list. Correlate the result list of the query with the previous list tag; then calculate the hit frequency by combining the number of visits and the number of hits, and finally adjust the order of the feedback results according to the above additional score.

(a) Hypothesis: number of document hits is $At_d$, number of user visits is $N_d$, so hit frequency can be expressed as:

$$f_d = \frac{At_d}{N_d} \tag{4}$$

(b) Define the similarity matrix as ‘sim’, the number of labels extracted by a document is vector $d_t$, and the default time period is vector $T_{cycle}$, so the matrix can be expressed as:

$$simSort = sim(f_d, d_t, T_{cycle}) \tag{5}$$

(c) The calculation of final rating is:

$$Search\_score = Wt_a \times simSort + Wt_b \times tf - idf_{i,j} \tag{6}$$

It is necessary to clarify that $Wt_x$ as weight value when calculating the final score. In order to avoid the number of visits and time period excessively affecting the final filtering result, the algorithm
adopts the cluster analysis method to calculate the heat and word frequency probability caused by the time period. After several trials, Sorting effect is most suitable when $W_{t_a}:W_{t_b} = 0.55:1.2$.

4. Research Results and Analysis

4.1 Test Design

The experiment is divided into two implementation steps: firstly, the distributed storage method is used for data storage, and the operation is continued in 5% increments every day that different fields of data are entered into the database, at the same time, each data is inverted indexed, then basic data pool is established. secondly, the original search tool of WinSearch is used to retrieve the original search data results, use the optimized algorithm to search and compare the same data volume.

In the 45-day follow-up monitoring, the experiment selected people of different ages and different hobbies to participate. The first step gives 20 task target sets of $T$ that need to be retrieved, the first step we will give 20 task target sets $<T>$ that need to be retrieved, extracts representative related statements within the target, then breaks the statements into separate words, finally it will be built into a search vocabulary set $<W>$. The second step, they retrieves each vocabulary to find the top 30 data built a set $<M>$ from set $<W>$, and obtains $<MX>$ according to the result set $<M>$ of each element combination in the set $<T>$, then attaches the current time stamp; Finally, the experimenter looped through the search behavior during in 45 days and get the data set $<Time-MX>$.

4.2 Analysis of Research Results and Discussion

The judgment of this result is based on the correlation between the search results of $<Time-MX>$ and WinSearch with query target set $<T>$. According to the degree of relevance of the content, the evaluation scores are divided into 1 to 5 levels, then the results of the ranking are evaluated by the experimenter, which is also divided into 5 score levels, and the experimental results are shown in following figures.

![Fig.1 Result score](image1)

![Fig.2 Data accuracy](image2)

Compared with the original search method, Compared with the original search method, the accuracy will decrease when the initial data increases, but the WinSearch algorithm only considers the matching vocabulary of the file and the relevance of the content, when the new data enters, it will continuously affect the user's actual retrieval target and ordering. The optimization algorithm proposed in this paper takes the user's retrieval content and the frequency of the hit data into the retrieval score for a period of time, thus ensuring the stability of the retrieval result, as increasing of the data correlation information, a satisfactory search sorted list is finally provided to the user.

5. Summary

This paper deals with data screening, feedback sorting, etc. for unstructured data. When considering behavior of user search and target intention, this paper give a type of information retrieval
optimization algorithm to solve those problem which based on heat computing and category labels in the period. After experimental comparison, the algorithm is superior to WinSearch algorithm. At the same time, the algorithm has higher sensitivity to time period and category attributes, so it is necessary to give manual assistance according to the actual situation. This is also the limitation of this algorithm and the place where it can be improved in the future. Therefore, the use of big data computing and intelligent analysis technology to adjust various attributes between different users and build a diverse data framework has become the next research direction.

References

[1]. Rui Zhang. An intelligent retrieval algorithm for enzyme ontology based semantic extension. [D], Shandong Normal University, 2015(06), p.24-29.

[2]. Jianfeng Yang. Function optimization problem based on genetic algorithm and ant algorithm. [J]. Journal of Zhejiang University (Engineering Edition). 2007 (03), p.428-430.

[3]. Shi Wang. (2016). research and Optimization on keyword retrieval algorithm for big data. [D], North China University of Technology,2016(03), p.34-41.

[4]. Qi Yang, ZhenCan Lin, Fan Huang, et al. Database Full-text Retrieval System Based on Hibernate Search [J]. Computer Engineering, 2010, 36(4), p.74-76.

[5]. ShaoNan Li. Research on fast retrieval Algorithm chinese expressions and sentences based on Chinese corpus[D]. Yanshan University,2013(12), p.21-26.

[6]. Chuanyao Yang, Yuqin,Li, Zhenghua, etc, A Yellow Page Information Retrieval System Based on Sorted Duality Inter-Relevant Successive Tree and Industry Ontology. Proceedings of 8th ACIS International Conference on Software Engineering Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD2007). 2007, p. 1147 – 1152.

[7]. DongLiang Chen, QingYuan Bai. Associated text classification based on word frequency vector [J]. Computer research and development, 2009, 46(z2), p.464-469.

[8]. Newstrom J, KeithDavis. Organizational behavior: human behavior at work / 9th ed[M], p.105-147.