Research on keyword indexing algorithm based on big data

Baofeng Hui
School of Physical and Electronic Information Engineering, Qinghai University Nationalties, Xining Qinghai, 810007, China
649053613@qq.com

Abstract. Keyword retrieval is widely used in various aspects of Web information, such as information processing, data mining, etc. Keyword search has been studied for a long time. Many research results have been obtained from keyword search of early relational database, keyword search of semi-structured data and keyword search of graph data. Given multiple keywords, the goal of keyword retrieval is to find the matching node of keywords and find the most compact data fragment containing all keywords. For traditional retrieval model exists in the retrieval of data redundancy, the disadvantage of fuzzy matching, retrieval information deviation, combined with the current hot, discusses the heterogeneous data integration and redundant data, efficient data classification, keyword retrieval model and the method, such as to big data environment, make full use of the traditional technology combined with spatial retrieval model and other technical storage model, improved classification algorithm, and optimize the retrieval algorithm, thus improve the operation efficiency of the algorithm, to provide users with a set of data storage, classification and retrieval in the integration of large data retrieval platform.

1. Introduction
Network is a virtual platform for information transmission, acceptance and sharing. It integrates all aspects of information to realize the sharing of different resources. In recent decades, with the exponential increase in the amount of information on the Internet, the use of the Internet to obtain information has gradually become the main way for most ordinary users to access to new information. Because the data volume increase sharply, user is difficult to through the understanding of the underlying data structure, and most users do not have related skills, as a result, users can get information by typing keywords retrieval rather than using a query language to retrieve, through set up in advance for big data platform for the mass information storage, classify the management of huge amounts of data, build up the necessary data model and data classification model, in a program can improve the efficiency of data retrieval and retrieval is more targeted.

Keyword search is the best choice for most users to use in perhaps network knowledge. Its advantage is: the user does not need to know to query the information of the underlying data structure, also don't need to learn any complicated query language, it only requires the user to provide the web search engine enter the keywords they are interested in, you can in the boundless data to find the information they want in the ocean [1-5]. This simple and convenient way to acquire knowledge is more and more favored by most network users.

The era of big data has led to great progress in social and economic activities. At the same time, the effective use of big data will bring inestimable value to big data. Massive electronic data, mobile Internet data, social media data, Internet of things sensor data, business data of enterprises and public institutions,
etc. are the data sources of big data today. If these data can be effectively managed, comprehensively analyzed, integrated and shared, huge social and economic benefits will be generated.

With the development of information retrieval technology, keyword retrieval is favored and accepted in text document and web page retrieval. For example, domestic search engine baidu and foreign Web search engine Google both adopt keyword search and query technology. Users can search for useful information in the vast data by typing keywords, combining current technology and Web service information search technology to filter and return a set of HTML documents matching keywords.

2. Foundation and technology of big data

2.1. The concept of big data

Big data refers to the acquisition, management and processing of data within a reasonable period of time, and the collation of data into a large amount of data for the purpose of positive business decisions. From these definitions, we can summarize the following conclusion: big data means that the existing hardware, software, algorithms and so on cannot provide enough support for the rapid growth of data, so we urgently need to develop new technologies to meet people's needs in time and space.

2.2. Big data generation

The generation of big data is the result of the wide application of computer network communication technology, especially the rapid development of the new generation of information technology such as Internet, mobile Internet, cloud computing, Internet of things and social network, which urges the rapid generation of data.

2.3. Big data storage and processing

2.3.1 Distributed storage and computing platform

Hadoop is a distributed file system and parallel execution environment platform, providing a convenient platform for processing massive data. It is characterized by the ability to store and manage petabyte level and above data. It is good at processing unstructured data and high-throughput data, and adopts the access mode of "one write, multiple reads".

Due to the distributed architecture design, Hadoop has high scalability and high fault tolerance, but it is not good at storing small data, a large number of random reads and needs to modify files and other occasions.

2.3.2 Distributed file system

Hadoop distributed File System (HDFS) is an open source implementation of GFS (Google File System), which runs on a cluster composed of several ordinary and cheap PC units. It hides the implementation details of load balancing and redundancy replication at the bottom and provides unified File System application program interface (API) to the upper layer. The HDFS architecture is shown in figure 1.

Fig 1. HDFS architecture

2.4. Big data analysis and visualization

2.4.1 Big data mining and analysis

(1) association analysis: also known as association mining, it is a relatively simple and practical analysis method, which can be used to dig out the correlation or correlation existing in a large number of data
sets, and then used to describe the relationship between a group of data items (rules and patterns). Common algorithms include, Apriori algorithm, FP-growth algorithm, etc.

(2) Classification analysis: a model is established for each category according to data characteristics, and data is allocated to different groups according to data attributes. Common algorithms include bayesian classification algorithm, nearest neighbor classification algorithm, decision tree classification algorithm, support vector machine classification algorithm, neural network, etc.

(3) Cluster analysis: on a similar basis, the data are divided into some different groups. The data in the same group are similar, while the data in different groups are greatly different. The difference between clustering analysis and classification analysis is that the classification of the former is unknown and needs to be determined according to the actual data. Common algorithms include k-means algorithm, nearest neighbor algorithm, neural network, etc.

(4) sequence analysis: it mainly analyzes a certain kind of time-related data in the data warehouse, traversal similar sequences or sub-sequences, mining time sequence patterns, cycles, trends and deviations, etc., which is a special correlation model (time attribute is added into the correlation model).

(5) Prediction model analysis: a prediction model is established by utilizing the distribution characteristics of certain attribute values of known data or objects in the database. Common methods include regression analysis, linear model, decision tree, genetic algorithm, etc.

2.4.2 Visualization technology
Visualization technology is divided into two categories: visual report and visual analysis. Visual reports use diagrams to describe, and are usually defined and characterized by information such as metrics and time, such as instrument dials, reports, dimensional analysis, and so on. Visual analysis enables users to have insight into data and make new discoveries, realize data interaction by using visual tools, enhance interactive statistical graphs by using point-and-click interface, and realize faster analysis, decision making, presentation and understanding of analysis results.

3. Classification model and algorithm
Classification is an important method of data mining, which can extract models describing important data classes from massive data to facilitate intelligent decision-making. Classification is to map data to a pre-defined class through a certain mapping relation, that is, to give a set of attribute vectors and their corresponding classes, and finally to obtain classification by using inductive learning algorithm [6].

Definition 1 given a data set \( D = \{d_1, d_2, \ldots, d_l\} \) and a set of category \( C = \{C_1, C_2, \ldots, C_k\} \)

The problem is to determine a mapping \( f: D \rightarrow C \), each mapping to a certain class, a class \( C_j \) contains mapping to this class

3.1. Naive bayes classification model
Naive bayes classification model is judged by evaluating the probability of each category relative to the data set based on the eigenvalues of the new data set, and bayes theory is adopted to determine [43].

\[
P(C_j|d) = \frac{P(C_j)p(d|C_j)}{P(d)}
\]  

(1)

In the formula, the denominator is not differentiated in each category and can be ignored. In the case of obtaining class variables, assuming that the characteristics are conditional independent, the formula can be simplified as:

\[
P(C_j|d) = P(C_j)\prod_{j=1}^{m}P(d|C_j)
\] 

(2)

The estimated value of \( P(C_j) \) \( \hat{P}(C_j) \) can be calculated by the percentage of training data.

\[
p(C = C_j) = \frac{N_j}{N}
\] 

(3)
In addition, \( P(d_i|c_j) \) estimate \( P^*(d_i|c_j) \) calculation is as follows:

\[
P^*(d_i|c_j) = \frac{1 + N_{ij}}{M + \sum_{k=1}^{N} N_{kj}}
\]

### 3.2. PCHA algorithm proposed

**Definition 1** (weighted Euclidean distance) \( x_{i_1}, x_{i_2}, ..., x_{i_m} \) and \( x_{j_1}, x_{j_2}, ..., x_{j_m} \) is an n-dimensional data object. If each variable is weighted according to its importance, \( w_1, w_2, ..., w_n \), then the weighted Euclidean distance can be expressed as:

\[
\text{dis}(x_i, x_j) = \sqrt{\sum_{k=1}^{n} w_k |x_{ik} - x_{jk}|}
\]

### 4. Unordered keyword retrieval algorithm

#### 4.1. DKRA algorithm proposed

**Definition 1** (keyword sequence) suppose \( k_1, k_2, ..., k_i \) is a set of retrieval keyword sequences and an abstract representation of data in big data, which can be represented as document keywords, video eigenvalues, audio eigenvalues, image eigenvalues, etc. This paper puts forward the expression in a broad sense. Narrow sense can be reduced to the traditional definition of the word, covering big data and small data [6-7].

**Definition 2** (data space) data space \( D \) refers to the multi-dimensional space formed by all keyword sequences in the entire data set. Any data item in the data set can be represented by vector dimension \( D \).

**Definition 3** (keyword weight) \( W_i \) is the keyword \( k_1, k_2, ..., k_i \) in data \( D_i \) is used to represent relevant data.

**Definition 4** (retrieval model) set of triples \( I = (D, Q, \varphi) \), the \( \varphi_j: Q_j \to 2^D \) for the first \( j \) query and the related data sets mapping, data sets, including \( D \) Q said query set.

**Definition 5** (retrieval optimization model), suppose the tuple \( M_j = (D_j, Q_j, \varphi_j) \), \( j = 1, 2, ..., n \). Is a partition of the \( J \)TH data retrieval model, \( \Phi: q_j: Q_j \to 2^D \) for the query and the mapping of the data set, several data sets, \( Q \), said the query set. Property 1 the more frequently the keyword items in the keyword retrieval sequence appear in the retrieval data, the greater the weight will be; otherwise, the smaller the weight will be.

#### 4.2. Algorithm verification and analysis

To verify the effectiveness of the algorithm, the following simple example is proposed:

- **Q**: "silver gold truck"
- **D1**: "shipment of gold damaged in a fire"
- **D2**: "delivery of silver arrived in a silver truck"
- **D3**: "shipment of gold arrived in a truck"

According to the above data, quantity \( d = 3 \). According to the formula \( idf_j = \log \left( \frac{d}{df_{j}} \right) \), the IDF value of each search term is as follows:

\[
\begin{align*}
idf_a &= 0, idf_m = 0, idf_a = \log(d|df_a) = \log(3/2) = 0.176 \\
idf_{af} &= 0, idf_d = \log(d|df_d) = \log(3/1) = 0.477, idf_s = 0.477 \\
idf_{sf} &= 0.477, idf_{sh} = 0.176, idf_f = 0.477, idf_t = 0.176, idf_g = 0.176
\end{align*}
\]

The formalized structure of keyword - data set vector is as follows:

Then the k-d matrix is constructed as follows:
Retrieval algorithm based on the analysis of the vector space model, the data space is put forward on the basis of space model to join the keyword sequence similarity calculation method with relevant data, and introduce the concept of K-D matrix, convenient data formal operation, a brief introduction to algorithm design analysis, finally through a simple example is given to illustrate the effectiveness of the algorithm, and a comparative analysis with other algorithms, manifests the superiority of DKRA algorithm, has obtained the good effect.

5. Ordered keyword retrieval algorithm

In today's era, it is necessary to treat data as a kind of resource and timely obtain useful information in massive data. Retrieval algorithm is an important means to promote the effective utilization of data resources. The essence of retrieval is to identify and match the retrieval keywords and data, so as to determine whether the retrieved data meet the requirements.

In the retrieval model, there is a set of keyword sequences, which are retrieved and returned to the relevant data set. This is the general process of keyword-based retrieval. A common retrieval model is vector space retrieval model, which introduces the concept of vectorization and constructs a set of keyword sequences into vectors. This model uses the Angle cosine or inner product between the retrieval keyword vectors and the vectors composed of data sets as the measurement, that is, the relevance of retrieval results. However, there is ambiguity between synonyms and synonyms in the model, as well as the error rate of the returned data set.

Aiming at the problem that the vector space retrieval model ignores the key word order combination, the key word sequence is ordered.

In this paper, the concept of position matching degree of search keywords is given, and the Orderly Keyword Retrieval Algorithm (OKRA) is proposed. The Algorithm can well solve the semantic differences caused by different sequences of Keyword sequences, so as to improve the original search algorithm and achieve the purpose of improving the quality of returned search data.

6. Forward and reverse indexes

6.1. Ranking index

Forward index (also known as forward index) is the premise and basis of inverted index. The fields are as follows:

1) Local ID field: represents the Local label of data.

2) Word ID field: represents the label after Word segmentation of data block, and is also the index Word label.

3) NHits field: represents the number of occurrences of an index term in the data block.

4) Hit List field: represents the position of an index word in the data block.

In essence, the index words are viewed by data label, that is, index words are retrieved by data label. Given any data label can be included in the index word, the number of index word, index word location, etc. The keyword index does not need to be retrieved by the data label, so the forward row index cannot meet the requirements. Even so, the forward row index is the basis and premise of creating inverted index, indispensable for calculating inverted index, and also the premise and basis of improving the algorithm.

The Local ID field of the positive index is numbered in ascending order (the number can be incremented by 1), which is for the convenience of calculation. In the process of inversion index transformation, because of the orderliness of Local ID in the forward index, the process of
transformation creation can ensure the orderliness of each word item in the corresponding data set of the inverted index. The forward index structure is shown in Table 1.

| LocalID | Key word ID | NHits | HitList |
|---------|-------------|-------|---------|
| D1      | K1          | m     | Hit 1, ..., Hit m |
|         | K2          | m     | Hit 1, ..., Hit m |
|         | ...         |       |         |
|         | Kn          |       |         |
| D2      | NULL        | ...   | ...     |
|         | NULL        |       |         |

6.2. OKRA algorithm proposed

6.2.1 Basic definitions
Definition 1 (retrieval step size) set \( p_1, p_2, ..., p_{\text{end}} \), is the corresponding position of a set of keyword retrieval sequences in the retrieval, then the retrieval step size is \( f_{ij} = p_j - p_i \).

Definition 2 (overall retrieval step size) set \( p_1, p_2, ..., p_{\text{end}} \), is the corresponding position of a set of keyword retrieval sequences in the retrieval, then the overall retrieval step size is \( f_{\text{all}} = p_{\text{end}} - 1 \).

Definition 3 (related data retrieval step size) set \( q_1, q_2, ..., q_{\text{end}} \), is the corresponding position of a set of keyword retrieval sequences in the relevant data, then the retrieval step size is \( g_{ij} = q_j - q_i \).

Definition 4 (overall correlation data query step size) set \( q_1, q_2, ..., q_{\text{end}} \), is the corresponding position of a set of keyword retrieval sequences in related data, then the overall retrieval step size of related data is \( g_{\text{all}} = q_{\text{end}} - q_1 (q_1 < q_2 < ... < q_n) \).

Property 1 the similarity between the search keyword sequence and the data in the data space is inversely proportional to the step length of two or more search keywords in the data, and is affected by the overall search step size (that is, the overall search step size is inversely proportional to the similarity).

Definition 5 (position matching degree) is set as \( g_{ij} \), is the step size of related data retrieval, and \( g_{\text{all}} \) is the step size of overall related data retrieval, then the calculation formula of position matching degree is

\[
poS_M = \frac{\sum_{i=1}^{n-1} g_{ij}}{g_{\text{all}}} * 100\%
\]

Algorithm verification and analysis
In order to verify the effectiveness of the algorithm, the following simple example is listed for verification:
The keyword retrieval sequence is Q: "truck silver gold", assuming that the following matches are obtained after indexing
Contents contained in the data:
D1: "truck silver gold damaged in a fire"
D2: "delivery of silver arrived in a truck"
D3: "shipment of gold arrived in a truck"
If the unordered keyword retrieval algorithm is used, all of the above data sets will be returned
Different sequences of key words lead to unsatisfactory results, which need to make use of the bits in the improved algorithm proposed in this paper
Set the match threshold to filter unqualified datasets and set the match threshold to threshold = 0
For data D1. Calculate \( g_{t,s} = 2 - 1 = 1 \), \( g_{s,g} = 3 - 2 = 1 \), \( g_{\text{all}} = 7 - 1 = 6 \)
For data D2, calculate $g_{ts} = 3 - 7 = -4, g_{all} = 7 - 1 = 6$ the position match $\text{posM} = \frac{-4}{6} \times 100\% = -66.67\%$, due to the $\text{posM} \leq 0$, incompliance requirements.

For data D3, calculate $g_{m,a} = 3 - 7 = -4, g_{all} = 7 - 1 = 6$ the position match $\text{posM} = \frac{-4}{6} \times 100\% = -66.67\%$, due to the $\text{posM} \leq 0$, is not in conformity with the requirements.

In view of the problems existing in the research of the disordered keyword retrieval algorithm mentioned above, the DKRA algorithm is further studied, and the ordered keyword retrieval algorithm is proposed. The purpose of this paper is to improve the accuracy of the retrieval data through the ordering of the key words, add the concept of position matching to the original algorithm, and use the position matching degree to affect the retrieval returned data in the calculation of similarity.

7. Conclusion

The main contribution of this paper is to research and propose data storage model, parallel classification hybrid algorithm, disordered keyword retrieval algorithm and ordered keyword retrieval algorithm based on keyword retrieval to return relevant data. Finally, on the basis of the disordered keyword retrieval algorithm, this paper introduces the concept of the ordering of the retrieval keywords, considers the position matching degree obtained by different sequences of the retrieval keywords while calculating the similarity, and proposes the ordered keyword retrieval algorithm and its improved form. The storage model proposed in this paper only provides a data storage technology solution, and the parallel classification hybrid model lacks research on parallel classification of heterogeneous data across nodes, and the proposed disordered keyword retrieval algorithm and ordered keyword retrieval algorithm need to further increase data set research and analysis.

References

[1] S. Agrawal, S. Chaudhuri and Cz Das, Dbxplorer: A system for keyword-based search over relational databases [A]. In Proceedings of the 18th International Conference on Data Engineering (ICDE) [C], 2002, 5-16.

[2] Cz Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti and S. Sudarshan. Keyword searching and browsing in databases using banks [A], In Proceedings of the 18th International Conference on Data Engineering (ICDE) [C], 2002, 431-440.

[3] V Hristidis, L. Gravano and Y Papakostantinou, Efficient ir-style keyword search over relational databases [A]. In Proceedings of the 29th International Conference on Very Large Data Bases (VLDB) [C], 2003, 850-861.

[4] V Hristidis and Y Papakostantinou, Discover: Keyword search in relational databases [A], In Proceedings of the 18th International Conference on Very Large Data Bases (VLDB) [C], 2002, 670-681.

[5] V Kacholia, S. Pundit, S. Chakrabarti, S. Sudarshan, R. Desai and H. Karambelkar. Bidirectional expansion for keyword search on graph databases [A], In Proceedings of the 31St International Conference on Very Large Data Bases (VLDB) [C], 2005, 505-516.

[6] Zhang sheng. A semantic retrieval model based on domain ontology [J]. Software guide, 2014, 13(03):18-20.

[7] Wang weiguo, xu weimin. A personalized query extension model based on latent semantic analysis [J]. Computer engineering, 2010,36(21):43-45. CAO H, BHARDWAJ A, GOVINDARAJU V. A probabilistic method for keyword retrieval in handwritten document images [J]. Pattern Recognition, 2009, 42(12):3374-3382.