Research on the Private Customized Information Retrieval based on Hadoop Cluster

Li Bo¹ ², Zhang Xiao¹ ² and Yan Jingyi ³

¹ School of Computer Science, Northwestern Polytechnical University, Xi’an Shaanxi 710129, China;
² Ministry of communications Key Laboratory of big data storage and management, Xi’an Shaanxi 710129, China;
³ School of Management, Northwestern Polytechnical University, Xi’an Shaanxi 710129, China

Corresponding author: e-mail: zhangxiao@nwpu.edu.cn

Abstract. With the development of science and technology, it is no longer a technical difficulty for users to find the answers they need from massive data of information retrieval. At present, with information retrieval methods, the most urgent need is how to find personalized search results that satisfy users and meet their needs. Therefore, private customized information retrieval technology has great research value. This paper builds a Hadoop distributed cluster search engine, improves the PageRank algorithm, adds classified attributes such as user tone, and further classifies and ranks the web pages. At the same time, the C4.5 algorithm is used to classify users, thus it provides private customized search service. Through comparative experimental analysis, it is proved that the improved method proposed in this paper is effective and has profound research value.

1. Introduction
With the development of science and technology, people are in the era of a big data explosion. According to the "37th Statistical Report on Internet Development in China" released by China Internet Network Information Center (CNNIC) in June 2016, the number of Chinese Internet users has reached 710 million at that time. At present, the digital content of the world is about 487 billion GB, which has created new research topics. How to search for satisfactory answers? Searching for answers is not difficult. The key is to get a "satisfactory" answer, which is suitable for the searcher's "private" order. When dealing with the current massive amount of data, even a single high-performance supercomputer will have a big burden. In order to improve the experimental accuracy and load balance, this paper chooses to build a Hadoop computer cluster. Hadoop is a very classic distributed computing platform, including the distributed file system HDFS (Hadoop Distributed File System) and the distributed computing framework MapReduce. By building a Hadoop cluster and optimizing search algorithms, semantic recognition, speech (tonal) recognition, generalized search, and providing customers with privately-tailored search services to optimize search results [1]-[2]-[3]-[4].

2. Research status
In the field of information retrieval, researchers began research in the mid-20th century. Many articles describes the information retrieval as follows: Information retrieval is a process in which users use...
some methods to search according to their own information needs. The current definition is to find information that meets user needs from a collection of large-scale unstructured data. In order to deal with massive amounts of data, search engines become an indispensable tool in people's lives. Currently, the search resources are no longer single text or words, and the users’ needs are more complex. The data format of information retrieval is more extensive. Voice search, image recognition, and multimedia resource-based search technology are also developed[5]. For example, Photo of Baidu is a typical image search engine.

Distributed computing is a commonly used method when dealing with massive data. Hadoop has the advantages of high scalability, high reliability, and low cost of use[6]. The research in this paper is based on Hadoop's distributed personalized retrieval. According to the user's behavior habits and the sentiment attitude when sending the search information, search engine will combine the classification algorithm C4.5 algorithm and provide the private customized search service, which best meet the user's desire.

3. Hadoop distributed file system
The retrieval structure of this paper adopts the architecture of distributed file system Hadoop to build a distributed cloud computing platform. Distributed computing is an effective solution to deal with big data problems, which important significance is each node handles huge data to achieve load balancing, resource sharing, and good scalability[7]. It is suitable for the retrieval requirements under the current big data era and can provide information retrieval.

3.1. Hadoop framework
Hadoop is currently the most widely used and most recognized parallel computing framework. It has two core components, HDFS and MapReduce. Due to the characteristics of Hadoop, users can concentrate on distributed upper-layer applications, do not care about how the distributed underlying layer is implemented, and make full use of the high performance of the cluster to achieve high-speed distributed computing.

3.1.1. Simple. Hadoop abstracts, modularize, and implements all of the underlying design details, which greatly increases development efficiency and makes development easier.

3.1.2. Fault tolerance. Hadoop's task allocation strategy can balance load for each node, so processing speed is fast and execution efficiency is high.

3.1.3. Efficient. Hadoop abstracts, modularize, and implements all of the underlying design details, which greatly increases development efficiency and makes development easier.

3.2. MapReduce concept
The current mainstream algorithm for information retrieval uses the PageRank algorithm. The PageRank algorithm mainly uses the MapReduce technology. MapReduce applies the ancient Chinese divide-and-conquer ideas to computer programming, which is inspired by traditional programming. For MapReduce, the common explanation is to compose a bunch of unstructured data into key-value pairs. Map parses these messy data extraction keys and values. The process of parsing is also called extracting the characteristics of these data. After the key-value pair is established, it will There is a Shuffle process. The essence of this Shuffle process is to clean up the memory, make the system more efficient, and internally sort these key-value pairs. The Reduce stage filters and organizes these key-value pairs according to the rules to get the required Data fragment. The working principle of MapReduce is shown in Figure 1.
Figure 1. Architecture diagram of Hadoop MapReduce

Hadoop MapReduce uses a master-slave architecture consisting of four core components: Client, JobTracker, TaskTracker, and Task. In MapReduce, the Client submits the job to the JobTracker and the program begins execution. JobTracker acts as the core controller in the execution of the job and acts as the commander. During the execution of the job, it monitors the status of each node. Once an exception occurs, it will take immediate action and assign it to other nodes for execution. The job was successfully completed. The TaskScheduler performs the JobTracker assigned tasks and periodically reports the execution status of the job to the JobTracker in the form of a heartbeat. The task is divided into MapTask and ReduceTask. MapTask parses the input information into key-value pairs to generate intermediate data. After the Shuffle and Merge processes, the sorted intermediate data is passed to the ReduceTask. ReduceTask organizes the intermediate results on demand and saves the results to HDFS for further processing.

3.3. HDFS Distributed File System

HDFS is an open source file system developed by researchers. The advantage of HDFS is that it is highly fault-tolerant, especially suitable for massive data sets, with low cost and reliable service. Figure 2 shows the architecture of HDFS.

Figure 2. Architecture diagram of HDFS

HDFS adopts the master-slave architecture mode, which has structural components such as NameNode, SecondaryNameNode, DataNode, and Block. The main task is to manage the metadata of HDFS and master the file directory tree of the entire HDFS and its directories and files. The NameNode is a control center that manages the namespace and controls the client's access to files to initiate monitoring. SecondaryNameNode
The metadata of the NameNode is backed up periodically, sometimes called checkpoint. The main purpose is to set up a special node to check and back up the metadata. The DataNode is responsible for saving data. All DataNodes perform file operations under the command of NameNode. HDFS is most suitable for processing large data sets. The data basically follows the rules of one-time write and multiple-read. The data on HDFS is saved in the form of a block on the DataNode.

4. Improvement of the search algorithm
In order to realize the “private” search service, the search algorithm is combined with the data mining algorithm.

4.1. Information Retrieval Algorithm
For search engines, the core search algorithm has a great impact on the performance of the platform. Voice input recognition and search is a technology that is now emerging as compared to traditional pure text search. At present, artificial intelligence is not good enough to integrate human emotions into data analysis. This paper introduces the measure of "tone" in the algorithm to carry out more accurate and humanized analysis. For the same sentence, the different tone may have different meanings or ideas. In addition to adding "tone" and user attitudes, this article also considers the starting point of the current search. Such as a user wants to search for related content, he maybe will exist, but if search engines still return he common responses and do not consider his tone, the user also is disappointed. According to the form of data mining, the search engine set the corresponding search according to the user's behavior and preferences, and applies the data mining method to the information retrieval, thereby providing the user with a private search service.

Information retrieval algorithm is the core component of the search engine. This paper uses the PageRank algorithm as search sorting algorithm and adds some new key points proposed in this paper. PageRank usually uses two hypotheses of quantity hypothesis and quality hypothesis to calculate. The simple PageRank algorithm flow can be expressed as a process in which a web page constitutes a directed graph. It is assumed that A, B, C, and D are the nodes of four web pages. If web page A has a link relationship with web page C, there is a directed edge $A \rightarrow C$, as Figure 3 show.

![Figure 3. Graph of PageRank algorithm](image)

If the user currently stays on the B page, the probability of the Internet user jumping to A, C, and D is $1/3$, and 3 means that B has 3 outgoing links. To abstract this problem, it can be represented by a convex combination of matrix $P$ and matrix $E$, which is defined as follows:

$$ G = \alpha P + (1 - \alpha)E $$

Here $\alpha < (0, 1)$ is the damping factor, the matrix $P$ is defined by the hyperlink structure of the network. $N$ is the dimension of the matrix $P$, and the PageRank problem is the eigenvector corresponding to the first eigenvalue 1 of the matrix $G$. The iterative tensor of the higher-order PageRank problem can be expressed as:

$$ \overline{M}(i, j, ..., l, k) = \alpha \overline{P}(i, j, ..., l, k) + (1 - \alpha)v_i $$

Figure 3. Graph of PageRank algorithm

4
Through the underlying PageRank algorithm, combined with the data mining and classification algorithms, the accurate information retrieval process of data analysis can be realized, what’s more, these also are the experiment of the "private” search information retrieval service.

4.2. Customer Classification Algorithm

In the data mining part of the overall algorithm, it is first necessary to analyze whether it is a "classification" or a "clustering" problem. In theory, each user is a separate cluster in search, they may have the same starting point in some specific searches, but in fact, each user’s needs may not be exactly the same, so the classification algorithm is more appropriate. In this paper, the C4.5 algorithm is used as the algorithm for classifying users. C4.5 constructs a classifier in the form of the decision tree processes a large number of data that needs to be classified and tries to predict the category to which the new data belongs.

4.2.1. Information Gain

The information gain is actually used in the ID3 algorithm to perform attribute selection metrics, and the attribute with the highest information gain is selected as the split attribute of the node N. The selected attributes minimize the amount of information required for tuple classification in the result partition. The expected information required to classify tuples in D is as follows:

\[ \text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \]  

(3)

Info (D) is also known as entropy. It is assumed that the tuple in D is divided according to attribute A, and attribute A divides D into v different classes. After this division, the information needed to obtain an accurate classification is measured by the following equation:

\[ \text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Info}(D_j) \]  

(4)

The information gain is defined as the difference between the original information requirement (ie, based on the class ratio only) and the new requirement (that is after the A is divided):

\[ \text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D) \]  

(5)

4.2.2. Information gain rate

In order to avoid the disadvantage of information gain, C4.5 behind ID3 adopts the concept of information gain rate. The information gain rate normalizes the information gain using the "split information” value. The classification information is similar to Info(D) and is defined as follows:

\[ \text{SplitInfo}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \]  

(6)

This value represents information generated by dividing the training data set D into v partitions corresponding to the v outputs of the attribute A test. Information gain rate definition:

\[ \text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}(A)} \]  

(7)

Through this improvement, the classification accuracy of C4.5 is improved and can be better used in user classification. Compared with other algorithms, C4.5 algorithm are easy to understand and have high accuracy. In this article, we use C4.5 as a classification algorithm to classify users and provide them with more humanized search results. The algorithm performs iterative update after each search by the user. If the user searches for data very frequently during this time period, it will perform iterative recalculation after the search peak is reduced. Through the combination of these two algorithms, it is possible to return accurate search results for a specific user.
5. Experiment
The design of this paper is verified by the Hadoop environment analysis in Section 3 of this paper, the algorithmic logic analysis in Section 4. The experiment runs on a cluster of 4 nodes, 1 master node, 3 slave nodes. Hadoop version is Hadoop-0.21.0, and JDK version is jdk1.6.0-21. Table 1 is the parameter configurations for each node.

Table 1. Experimental configuration

| Parameter     | Master node | Slave node |
|---------------|-------------|------------|
| cpu           | Intel Core Processor | Intel Pentium Processor |
| cpu GHz       | 2.20GHz     | 2.66GHz    |
| amount of cpu | 2           | 1          |
| memory        | 16G         | 8G         |
| hard disk     | 2T          | 1.5T       |

The search engine based on distributed clustering and unoptimized PageRank algorithm, the search engine based on a distributed cluster and PageRank algorithm optimized by this paper are tested and analyzed according to network data. The experimental process is divided into three stages: web crawling, web page analysis, and web page query. For the PageRank algorithm optimized by this paper, the process of classifying customers is also included.

5.1. Web crawling
To give users the most comprehensive information, search engines first need to obtain a large number of web pages, based on the download page, analysis, download, and analysis. By starting from the search engine's seed site, and constantly crawling the link pages, search engines will achieve web crawling.

5.2. Web page analysis
Search engines call the calculation module to calculate the topic of the web page, the degree of mirroring, the degree of entry, the degree of ranking. The core keywords of the web page will be analyzed. For the search engine, the originality of the website is very important. You can calculate which pages of each page of a website are linked to, and which pages link to the website so that each page has a weight value. In this process, the PageRank algorithm proposed in this paper adds classification factors such as "tone". The web crawling classification will be more refined and will be based on the searcher's preference.

5.3. Web page query
In this process, the role of the PageRank algorithm is reflected, and the improvements studied in this paper are mainly reflected in this stage. According to the search cache mechanism, the webpage has been downloaded and analyzed in the early stage, and then each webpage is indexed. In this process, the unimproved and improved PageRank algorithm will provide customers with different search results. We will analyze the pros and cons of the algorithm according to the recommended order of users clicking on the webpage. Remove the ad page. If the user clicks on the lower ranked page and clicks on different searches to get more pages or even research, the algorithm has lower recommendation accuracy. In this paper, the first result weight of the search is set to 1, and the order is incremented. If the search is researched, 10 is added, and the clock frequency of the user each time the search is completed is obtained, thereby obtaining the experimental result.
When the sample space is small, the traditional PageRank algorithm does not have too much difference with the improved PageRank algorithm. However, as the number of data increases, the improved PageRank algorithm will be greatly improved. The improved PageRank algorithm is more traditional. The PageRank algorithm has an obvious improvement. Overall, the algorithm proposed in this paper is effective and has great development value.

6. Conclusion
By constructing the experimental environment, simulating the actual experimental platform, conducting scientific verification and analysis, and drawing the experimental conclusions, through the addition of some new indicators, the improved PageRank algorithm has a great effect. By this method, the user's search accuracy, and satisfaction are higher, which proves the paper's improvement is effective, the idea is correct. Following works will be done in the future:

- The essential purpose of information retrieval is to give users a quick presentation of the data they want. No algorithm can adapt to the current environment 100%. Combining the attitudes of speech, tone, and speaker, this is very useful for human beings, but it still takes time to develop for computers.
- Most of the current information retrieval is the search for short or one-sided data, and there is no continuous data retrieval for continuous data. For example, the ability of a computer to read an article's paragraphs, or to make judgments on continuous searches, and the ability to logically analyze, this article will continue this module for analysis.
- The current SSD is a very common storage structure and is suitable for the case where the number of writes is small and the number of reads is large. For this type of storage device, this article will try to integrate SSD technology into the Hadoop file system to speed up the reading and give the user a better experience. However, the author believes that the prospect of SSD development is not particularly optimistic. The traditional disk system as the storage structure may be more stable and will ensure the stability of the data.

Acknowledgments
This work is supported by the MIIT Key Laboratory of Big Data Storage and Management, the NFS of China under Grant No.61502392 and No.61272323 and the Ministry of Science and Technology of China, National Key Research and Development Program (No.2018YFB1004401).

References
[1] Azzedin, Farag and Muthucumaru Maheswaran. “Towards Trust-Aware Resource Management in Grid Computing Systems.” 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID'02) (2002): 452-452.
[2] Sun Yongchao. Research on information retrieval system based on Hadoop [J]. intelligence exploration, 2016, 1 (8): 125-130.
[3] Lee, Kyong-Ha et al. “Parallel data processing with MapReduce: a survey.” SIGMOD Record 40 (2011): 11-20.
[4] Li Bo, Zhu Chenqi, Liu Jiamin, He Jing, Ma Tiao. Wireless Charging Device Based on Magnetic Resonance Technology [J]. Electronic Design Engineering, 2016, 24 (10): 111-113.
[5] Beloglazov, Anton and Rajkumar Buyya. “Energy Efficient Resource Management in Virtualized Cloud Data Centers.” 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing (2010): 826-831.
[6] Wang Peng. Research and application based on Hadoop full-text retrieval [D]. Chongqing University of Posts and Telecommunications, 2017.
[7] Schwarzkopf, Malte et al. “Omega: flexible, scalable schedulers for large compute clusters.” EuroSys (2013).