The Research on Information Collection and Retrieval Technology of Big Data based Cloud Computing

Hongjun CHEN
Computer science and Technology Specialty of Sichuan TOP Vocational Institute
Chengdu, China
e-mail:2009missxiaochen@163.com

Kun HUANG
Computer science and Technology Specialty of Sichuan TOP Vocational Institute
Chengdu, China
e-mail: huangkun8085@sohu.com

***Abstract*— with the rapid development of cloud computing, information shows explosive growth. The Cheap cloud storage and computing power, also contributed to the generation and applications of big data. The big data is unstructured data more than 50%, so much of them are stored as files in the file system. The big data is divided into many parts that stored into chunk server, and generates the corresponding metadata that stored into the master server. Then how to collect the web-url and the terms, and how to retrieval is be researched.**

Keywords- cloud computing ;big data;chunk;server; collect; retrieval; collection;web

I. INTRODUCTION

With the rapid development of cloud services, more and more individuals and businesses will migrate to online business, in order to reduce hardware and system maintenance costs. Online trading, online social networking, automatic sensors, mobile devices, and scientific instruments generated a lot of data. In addition to those fixed data production source, a variety of trading practices may also accelerate the accumulation of data speeds. For example, the social multimedia data grows explosively, which was derived from online transactions and new record behavior. The data is huge which maybe is presented by GB、TB、PB，and EB, even ZB. Data is always in growth. Some of them are big data. What's the big data? Can cloud computing handle it right? How to get the useful information form them? If the retrieved and useful is necessary for everyone's everyday life, they facilitate greatly the entire human society undoubtedly. And then the cloud will show the great value. Therefore, the research on information collection and retrieval of big data in cloud computing will become a hot topic.

II. BIG DATA STORAGE CHARACTERISTICS AND METADATA

Big Data is generated with super storage and computing power of cloud computing。The so-called big data refers to large amounts of unstructured and semi-structured data. It has four characteristics. First, a huge amount of data volume (Volume), the second is the variety data type (Variety), three is the low value density (Value), and four is the high processing speed (Velocity). Such characteristics of big data show that their storage、collection and retrieval aredifferent from the traditional ways. It only depends on the cloud's abilities to function.

If save the big data into a relational database for analysis ， you will spend too much time and money. And the more than 50% of big data is unstructured data. So more of them are stored as files in the file system. Currently, cloud-based cluster file system has become the main carrier of the large-scale data centers. Now, in cloud computing the file system storage technologies are mainly GFS, HDFS. Below GFS [1] indicates that how the big data is stored. See Figure 1.

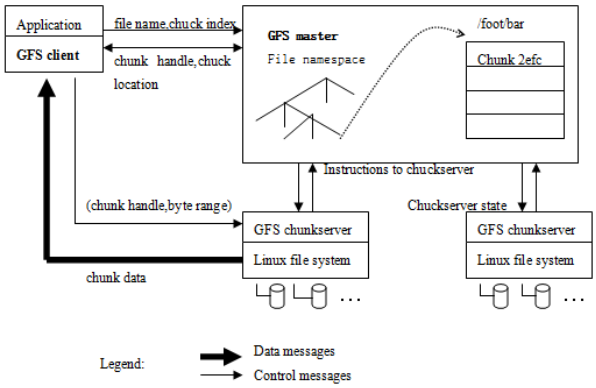

Figure 1. Big data stored in the GFS

Cluster file system, the basic idea is: a big data file is divided into many chunks; the chunk data is stored in the chunk server, each chunk has the corresponding metadata. Metadata is stored in the master server. The master server saves the three main types of metadata: namespace of files and chunks; the mapping file to the chunk; each chunk copy location. All metadata are stored in the master server's memory. Namespace metadata is used to maintain the file system name space, by querying the name space to query the attribute information of the specified file or directory path which can display contents that the directory contains. The other metadata records the storage location of the file, by which you can query and know the file offset file data chunk address. In the form of keywords metadata are organized into dynamically scalable index structure[2]. If the text information presented not directly can be misused, and can mislead search engines. So Google pay more attention to metadata.

III. TWO MAIN PROCESSES

Value density is inversely proportional to the total amount of data. In the hundreds of billions of web pages, the ones that truly meet their needs may be from several hundreds to thousands in billions pages. For example, a 1-hour video, uninterrupted monitoring, useful data may be only 1 to 2 seconds. In the strong technical capabilities support of cloud computing, how to purify the valuable information from millions and millions of cloud servers becomes urgent problem. In the aspect of information retrieval, Baidu and Google are the leader. In the cloud-based era of big data, some of core idea is still draws from Google search. Cloud-based big data information retrieval is divided into two main processes: Web collection process and the retrieval process.

IV. WEB COLLECTION

Retrieve the metadata in the master servers in the cloud, and analyze to sort the metadata. Then draw the he top correlation chunk server that will be made parallel and distributed search. Retrieval methods are mainly "breadth-first", "depth-first." Retrieval results are collected in the Index Repository [3]. The user request of search is really the search in Index Repository. The combination of web content and index pages is computed by PageRank algorithm in order to get the inverted list that will be still stored in Index Repository. Page title and link data are stored in an index for breadth-first search; web content is stored in another index in order to retrieve low- frequency long tail、personalized, depth-first search. See Figure 2.

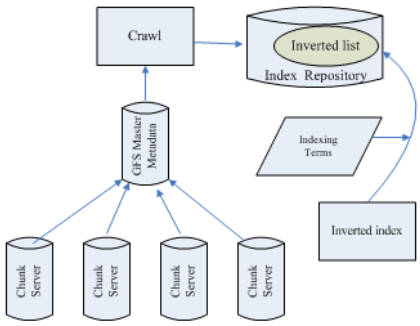

Figure 2. Web collection process of big data

In the process of whether web collection or the search requested by the user, the correlation calculation method will be used[4]. The parameters of Correlation calculation method are described. N represents the total number of keywords; M represents the total number of all the online resources; T = {t1, t2, ..., tN} represents the indexed keywords set; R = {r1, r2, ..., rN} represents the set of index metadata resource. TT = {t1t1, t1t2,…, tNtN} represents correlation between keywords; TR = {t1r1, t1r2, ..., tNrN} represents correlation between the keyword and the resource; FT: keyword use frequency; FR: GFS metadata frequency of use; Ri represents online resource set marked by tikeyword; Rj represents online resource set marked by tj keyword[5].

The method to calculate correlation t_it_j between keyword t_i and keyword t_i:

$$t_it_j = \frac{|R_i \cap R_j|}{|R_i \cup R_j|} \quad (1)$$

The method to calculate correlation t_ir_j between keyword t_i and online resources r_i:

$$t_ir_j = \frac{c_{ij}}{FR_j} * \log\left(\frac{M}{|R_i|}\right) \quad (2)$$

User query results obtained is not timely, but the results is ranked in the search engine's cache area, of course, search engines do not know the future, he will not know which keywords users will query. But a keyword thesaurus was built, and when processing user queries, word segmentation will be first conducted in accordance with the vocabulary word in order to get the similar keywords. So search engines can index by the keywords from the inverted sorted-list in Index Repository [3].

How to do URL ranking? PageRank algorithm is still be used. The basic idea of PageRank is that: If a network is pointing to other pages many times, which indicates that this page is more important or high quality. In addition to considering the number of Web links outside, the links page itself levels is also considered, and the number of forward links on this page to other pages. Higher weight pages have high level. Simplified formula of PageRank [4]:

$$\Pi P(A) = (1-\delta) + \delta(\Pi P(T1)/X(T1) + ... + \Pi P(T\iota)/X(T\iota)) \quad (\iota=1,..., N) \quad (3)$$

Description of each parameter: PR (A): PageRank of web page A; PR (Ti): web page's Ti PageRank link to page A; C (Ti): Ti outbound links page number; d: damping.

V. INFORMATION RETRIEVAL

When a use query, by user asking questions, the user requests are submitted to the search- agent. Search- agent sends the question to Index Repository. And re-sort the search results by correlation, then send the sorted result to the display web presented to the user. So improve greatly search capabilities and retrieval speed. See Figure 3.

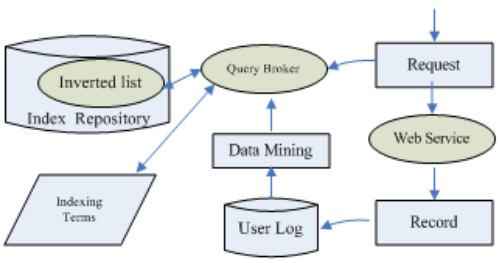

Figure 3. Information Retrieval of big data

A. *Analysis of the submitted search requests*

Search user inputs questions called keywords in the search engine, submit search requests. The search engine will search for the request for detailed analysis [5].How to make word segmentation. English request word segmentation is simple because each word separated by a space. It is easy to get words set. After eliminating duplicates, the query keywords are gotten. But Chinese request word segmentation is more complicated, and it is also important and necessary for the Chinese users. Chinese word segmentation has three common methods as the following:

1) Matching based on string

Matching based on string has three methods [6]: forward maximum matching、reverse maximum matching、at least split. Forward maximum matching method is that from left to right to segment the question to some words. Reverse maximum matching method is that from right to left to segment the question to some words. Minimum segmentation: Cut out the required number of words is the least.

2) Understanding segmentation method

Search engine simulates human thinking to understanding its statement for segmentation technology. Words and expressions segmentation will integrate for understand. The basic idea is that segment and handles ambiguity by syntactic and semantic analysis. It usually consists of three parts: the segmentation subsystem, syntactic and semantic subsystem, the total control section. Under the coordination of the total control section, the segmentation subsystem can get information about words, sentences, and other syntactic semantic information. It simulates the human process of understanding the sentence.

3) Statistical segmentation method.

More times of occurrence of the neighboring word, Chinese word segmentation is more likely segment the word to a keyword.

Note that Remove stop words. When a user input the question request, more or less there will be a lot of stop words like "the", "you", “one of ”,”many”,” maybe”. When make word segmentation, they will be removed.

B. *Matching search request*

When a search engine to search requests received after detailed analysis, it will match the index terms form index terms library. The terms include such as URL 、title abstract pieces etc. After matching thousands of URL, the correlation between request keywords and index terms will be sorted.

C. *Sort search results*

When make the web collection， the Inverted list has been in accordance with the PageRank algorithm. The number of the matched result is from hundreds to thousands. So Web search engines to filter the matching url. In the process, in addition to considering Inverted list, there are other factors. The searcher will combined PageRank value with the words document and the linked file description pages, Such as title, bold, bold, links, etc, even together with the search engine business advertisement [7] .

VI. TEST

The test environment configuration: 1 master server, 2 copies of the master server, 17 chunk servers, 1 Index Repository, 1 Index agent, 20 clients. 2 switches connected with 1Gbps line. 20 servers connected to a switch. All machine configuration: Intel Core™ i3@2.30GHZ CPU,4.0GB main memory, two 400G 5400rpm hard drives, 100 Mbps full-duplex Ethernet connection to the HP2524 switch. Egothor, a full-text search engine ,which is Open, efficient and cross-platform, is configured in Index agent.

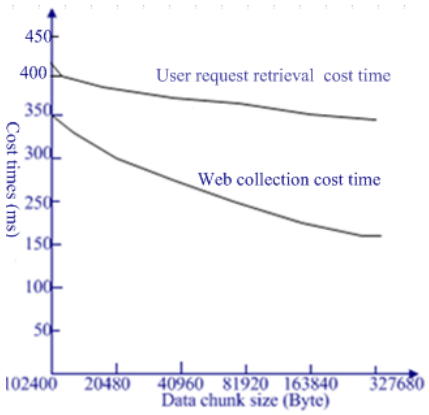

Figure 4. Relationships graph of chunk size and cost time

Be seen, with the rapid increase of chunk, web collection time decreased more slowly, user request retrieval cost time too. So the appropriate chunk size can improve the efficiency of web collection and retrieval. And with the servers increase, the computing capacity is stronger, the speed will be higher.

VII. SUMMARY

Big data retrieval technology still faces many challenges. Such as how to effectively deal with media types data retrieval such as graphics, sound, video etc.? Growing popularity of intelligent terminals, how to make use of information retrieval, even information recommended in the intelligent terminal? In short, the large data retrieval technology research still has a long way to go.

REFERENCES

[1] Hongjun Chen, Research of Cloud Storage and Data Read-write Technology, Applied Mechanics and Materials Vols. 347-350 (2013) pp 3555-3559

[2] Google File System (GFS), http://wenku.baidu.com/view/8a839535ee06eff9aef8074d.html,2012.

[3] Web search engines process principle and crawl and alysis,http://idc.cnw.com.cn/SEO/htm2009/20091123_186516_2.shtml, 2012

[4] PageRank,http://www.blueidea.com/tech/site/2003/1482.asp

[5] Google search engine, http://wenku.baidu.com/view/084dde1e650e52ea55189808.html

[6] LIU Bin, YANG Fan, Support multidimensional tag cloud mobile restaurant recommendation simulation system 2012，48（4）

[7] Google search engine,
http://www.599web.com/html/2012/sseo_0513/255.html