# Massive Medical Images Retrieval System Based on Hadoop

YAO Qing-An [1], ZHENG Hong [1], XU Zhong-Yu [1], WU Qiong [2], LI Zi-Wei [2], and Yun Lifen [3]
1. College of Computer Science and Engineering, Changchun University of Technology, Changchun, China
2. College of Humanities and Information, Changchun University of Technology, Changchun, China
3. Mississippi State University, Civil and Environmental Engineering, Mississippi State, USA

***Abstract*—In order to improve the efficiency of massive medical images retrieval, against the defects of the single-node medical image retrieval system, a massive medical images retrieval system based on Hadoop is put forward. Brushlet transform and Local binary patterns algorithm are introduced firstly to extract characteristics of the medical example image, and store the image feature library in the HDFS. Then using the Map to match the example image features with the features in the feature library, while the Reduce to receive the calculation results of each Map task and ranking the results according to the size of the similarity. At the end, find the optimal retrieval results of the medical images according to the ranking results. The experimental results show that compared with other medical image retrieval systems, the Hadoop based medical image retrieval system can reduce the time of image storage and retrieval, and improve the image retrieval speed.**

***Index Terms*—Medical Image Retrieval; Feature Library; Brushlet Transform; Local Binary Patterns; Distributed System**

## I. INTRODUCTION

The development of digital sensor technology and storage device leads to the rapid expansion of the digital image library, and all kinds of digital equipment produce vast amounts of images every day. So how to effectively organize the management and access of these images becomes a hot research direction in recent years. The traditional text-based image retrieval system uses the key words to retrieve the marked images. But owing to the limitations that artificial marking causes large workload, the content of the images cannot be completely described by words, and the understanding of images is different from person to person and so on, the text-based image retrieval system cannot meet the requirements for massive images retrieval. And how to carry on the effective management and organization of these medical images to provide services to clinical diagnosis becomes a problem faced by medical workers [1]. The content-based medical image retrieval (CBMIR) has the advantages of high retrieval speed and high precision and so on, and has been widely applied in the fields such as medical teaching, aided medical diagnosing, medical information management, etc [2].

The content-based image retrieval [3] is a kind of technology which makes use of the visual features of the images to carry on the image retrieval. Under the premise of a given query image, and according to the information of the image content or the query standard, it searches and finds out the images that meet the query requirements in the image library. There are mainly three key steps: first, selecting the appropriate image characteristics; second, adopting the effective feature extraction method; third, using the effective feature matching algorithm. Features that can be extracted from an image include color, texture, shape, flat space corresponding relation, etc. Color can be presented by color moment, histogram, etc. Texture can extract the Tamura feature, Gabor and wavelet transform of image. Shape can be divided into area-based method and edge-based method. Flat space corresponding relation can be described through two-dimensional string [4].

At present, many institutions have further studied CBMIR, and developed systems that went into practice. Such as the earliest commercial QBIC system [5] developed by IBM, WebSeek system [6] by Columbia University, Photobook system [7] by Massachusetts Institute of Technology and so on. There are also many outstanding works in the content-based image retrieval direction in recent years, for example, literature [8], based on the clustering of unsupervised learning, are the typical examples of CBMIR technology, literature [8][9] use the semi-supervised learning method, literature [9] carry on image retrieval with the method of relevance feedback, and a lot of works also improve the quality of image retrieval by improving the method of feature extraction, such as literature [11, 12]. The CBMIR algorithm needs to calculate the similarity between the features of sample medical images and the features in the feature library. It is a typical data-intensive computing process [13]. When the number of the features in the library is large, the efficiency of the single-node retrieval in the traditional browser/server mode (B/S) is difficult to meet the real-time requirements of the images, and the system has a poor stability and extensibility [14]. Cloud computing can assign the tasks to each work node to complete the tasks together, and with a distributed and parallel processing ability, it provides a new research idea for medical image retrieval [15].


doi:10.4304/jmm.9.2.216-222

Hadoop is an open-source project under the Apache Software Foundation organization, which provides the reliable and scalable software under distributed computing environment. It is a framework that allows users easily using and distributing computing platform, and it can support thousands of computing with PB-level nodes data [11, 12]. Hadoop distributed computing platform is suitable for all kinds of resources, data and other deployed on inexpensive machines for distributed storage and distributed management. It is with high reliability, scalability, efficiency and high fault tolerance, etc., and it can effectively improve the image the speed of retrieval. The text on the basis of reaching open-source framework Hadoop, analyzing the traditional image retrieval system and combining the content-based image retrieval technology and MapReduce computing framework [13] stores the image feature in database HDFS and developers the realized Hadoop-based mass image Retrieval System.

Hadoop Distributed File System (HDFS) is a scalable distributed file system. For it can be run on a cheap and ordinary hardware, it is supported by many companies, such as Google, Amazon, Yahoo! and so on. Under the circumstance that the underlying details is unknown, using the Map/Reduce functions to realize the parallel computing easily has been widely applied in the field of mass data processing [16]. Making use of the advantages of Hadoop, the problem that the retrieval efficiency is low in the process of medical image retrieval can be better solved, and there is no related research in the domestic presently [17]. Content-based image retrieval CBIR is the underlying objective retrieval by using global and local features of the image. Global features include color, shape, texture and so on; local features include SIFT, PCA-SIFT, SURF and so on [14]. As an automatic objective reflection image content-based retrieval method, CBIR is suitable for mass image retrieval. Semantic retrieval is the direction of development of CBIR image, but the image semantic has the characteristics of complexity, subjectivity, etc., and it is difficult in the extraction, expression and application of technical exist [15]. There are two main aspects of development of parallel image processing system; one is for some algorithms. It is searching the efficient parallel algorithm and development of high-performance parallel computer to achieve specific purposes, but such system is limited to the scope of application. The other is developed for general-purpose parallel image processing system, which is the mainstream of the parallel image processing system [16]. Image parallel computing generally are divided into two kinds: pipelined parallel and data parallel. Pipelined parallel is with the handling unit sequentially connected in series, that is, the output of a processing unit and the input of the next processing unit is connected. Data parallelism is composed of a plurality of processing units in parallel arrays, and each processing unit can perform its tasks independently [17]. With the increasing of the image data, the mass of the image retrieval process has become a very time consuming process.

To improve the efficiency of medical image retrieval, aiming at the shortage of the B/S single-node system, a medical image retrieval system based on the distributed Hadoop is put forward. And the experimental results show that the Hadoop-based medical image retrieval system not only reduces the time of image retrieval, improves the efficiency of image retrieval, but also presents a more apparent advantage for massive medical images retrieval.

Main innovations of this paper:

(a) With the continuous development of digital technology, there is a sharp increase in the amount of image data for the image data. For the mass interested in image retrieval problem of low efficiency, as well as the deficiencies of B / S single-node system, the efficiency of medical image retrieval is further improved and Medical Image Retrieval system based on Hadoop Distributed is proposed. It is based on Hadoop cloud computing platform, adopts the parallel retrieval technology and uses the SIFT Scale Invariant Feature Transform algorithm to solve the problem of massive image retrieval.

(b) Medical Image Retrieval system based on the Hadoop Distributed improves the efficiency of image storage and retrieval, which get better search results. They are mainly showing in the following aspects: medical image retrieval can meet real-time requirements of medical image retrieval, especially when dealing with large-scale medical image. It has the unparalleled advantages compared to traditional B / S single-node, and at the same time it reduces the image retrieval time and improves the efficiency of image retrieval, especially for massive medical image retrieval.

## II. HADOOP DISTRIBUTED MEDICAL IMAGE RETRIEVAL

### *A. Hadoop Platform*

Hadoop platform is the most widely used open source cloud computing programming platform nowadays. It is an open source framework which runs large database to deal with application programs on the cluster, and it supports the use of MapReduce distributed scheduling model to implement the virtualization management, scheduling and sharing of resources [10].

The structure of HDFS is that a HDFS cluster consists of a master server (NameNode) and multiple chunk servers (DataNode), and accessed by multiple clients. The NameNode is responsible for managing the namespace of the file system and the access of the clients to the files, while DataNode manages the storage of the data of its node, handles the client's reading and writing requests of the file system, as well as carries on the creation, deletion and copy of the data block under the unified scheduling of NameNode [11]. HDFS cuts the files into pieces, then stores them in different DataNode dispersedly, and each piece can be copied and stored in different DataNode. Therefore, HDFS has high fault tolerance and high throughout of data reading and writing.

MapReduce is a programming model, which is used for the calculation of large amount of data. For the calculation of large amount of data, the usually adopted

processing technique is parallel computing. First of all, breaking a logically complete larger task into subtasks, then according to the information of the tasks, using appropriate strategies, the system assigns the different tasks to different resource nodes for their running. When all the subtasks have been finished, the processing of the whole large task is finished. Finally, the processing result is sent to the user [12]. In the Map phase, each Map task calculates the data assigned, and then maps the result data to the corresponding Reduce task, according to the key value output by Map.

In the Reduce phase, each Reduce task carries on the further gathering processing of the data received and obtains the output results. To make the data processing cycle of MapReduce more visual, the calculation process of the MapReduce model is shown in Figure 1.

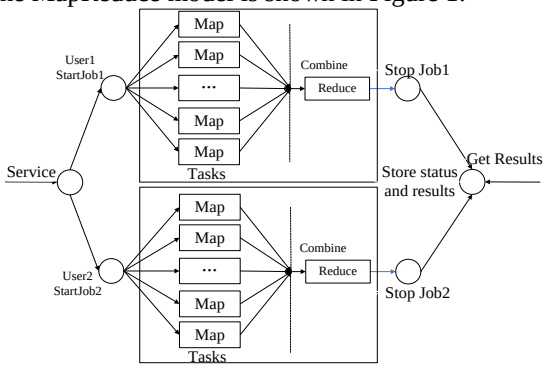


Figure 1. Data processing cycle of map reduce

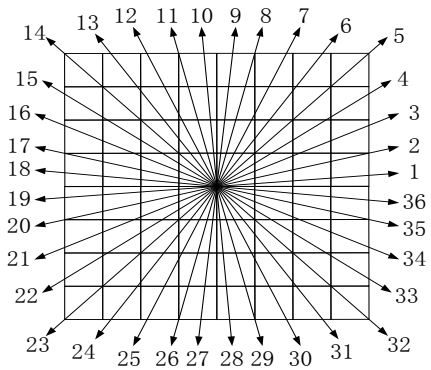


Figure 2. Level three decomposition direction of brushlet

### *B. Feature Extraction of Brushlet Domain*

Brushlet transform is the image multi-scale geometric analysis tool which aims at solving the problem of angular resolution. The two-dimensional Brushlets has certain direction structure and vibration frequency range, and can be reconstructed perfectly. The structure size of its basic function is inversely proportional to the size of the analysis window. The two-dimensional Brushlet with phase parameters shows its direction, thus better reflects the direction information of the image, and can conduct the decomposition of the Fourier domain [13]. Level one of Brushlet transform will divide the Fourier plane into four quadrants, and the coefficient is divided into four sub-bands, the corresponding direction is $\pi/4+k\pi/2$, $k=0,1,2,3$. Level two further divides each quadrant into four parts on the basis of level one, and the whole twelve directions respectively are $\pi/12+k\pi/6$, $k=0,1,\cdots 11$. There are sixteen coefficients after the decomposition, among which the four sub-bands around the center are with low frequency component, and the rest are with high frequency component. And so on in a similar fashion. Figure 2 is the decomposition direction graph of level three.

Given an image $f$, and conducts level $l$ decomposition of Brushlet to it, there are will be two parts after the decomposition, which are the real part $\hat{f}_r$ and the imaginary part $\hat{f}_i$. Each part has $4^l$ sub-bands, and each sub-band reflects the direction information of its corresponding decomposition direction. The place where the energy focused is exactly the parts where the texture image mutations. For each sub-band, its energy information can choose to be shown by the mean value and the standard deviation of the module value. Because Brushlet is a complex function, the corresponding sub-band coefficient of the real part and the imaginary part after the decomposition is used to calculate the module value at the same time. After the decomposition, the $n$ sub-bands of the real part and the imaginary part respectively are marked to be $\hat{f}_{nr}$ and $\hat{f}_{ni}$ $n=1,2,\cdots 4^l$. The mean value $\mu_n$ and the standard deviation $\sigma_n$ of the $n$ sub-band's module value respectively are:

$$\begin{aligned}\mu_n &= \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}|\hat{f}_n(i,j)| \\ &= \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\sqrt{[\hat{f}_{nr}(i,j)]^2+[\hat{f}_{ni}(i,j)]^2}\end{aligned} \tag{1}$$

$$\sigma_n = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(|\hat{f}_n(i,j)|-\mu_n)^2} \tag{2}$$

In the above equation, $i=1,2,\cdots,M$ , $j=1,2,\cdots,N$. M and N respectively represents the line number and the column number of each sub-band. The feature vector of image $f$ is:

$$F=[\mu_1,\sigma_1,\mu_2,\sigma_2,\cdots] \tag{3}$$

### *C. Feature Extraction of LBP*

LBP can depict the changes relative to the center of the pixel's gray level within the territory. It pays attention to the changes of the pixel's gray level, which is in accordance with human's visual perception features of the image, and the histogram is treated as the airspace characteristics of the image.

$$LBP_3^{u2}=\begin{cases}\sum_{i=0}^{7}s(g_i-g_c)2^i, & U(LBP_3)\le 2\\ 256, & \text{otherwise}\end{cases} \tag{4}$$

Among which:

$$s(g_i - g_c) = \begin{cases} 1, & g_i - g_c \geq 0 \\ 0, & g_i - g_c < 0 \end{cases} \quad (5)$$

$$U(LBP_3) = |s(g_7 - g_c) - s(g_0 - g_c)| + \sum_{i=1}^{7} |s(g_i - g_c) - s(g_{i-1} - g_c)| \quad (6)$$

In the above equation, $g_c$ is the pixel's gray value of a neighborhood center, and $g_i$ means each pixel's gray value of the neighborhood in clockwise, which is within the range of 3×3, and $g_c$ as the center.

*D. The Similarity Matching*

To measure the feature similarity of Brushlet domain, the average distance is used:

$$Sim_{\text{Brushlet}}(P,Q) = \sum_{i=1}^{6} |E_{Pi} - E_{Qi}| \quad (7)$$

Among which, P is the medical image waiting to be retrieved, and Q is the image of the medical image library.

For the LBP features of the image, firstly the characteristics are being unanimously processed, and then the Euclidean distance is used to calculate the similarity.

$$Sim_{\text{LBP}}(P,Q) = \sqrt{\sum_{i=1}^{32} \left(\overline{\mathbf{W}}_{Pi} - \overline{\mathbf{W}}_{Qi}\right)^2} \quad (8)$$

In the above equation, $\overline{\mathbf{W}}$ represents the characteristic vector after the normalization.

Because the value range of $Sim'_{\text{Brushlet}}$ and $Sim_{\text{LBP}}$ is different, the external normalization is being processed to them. The specific process is as follows:

$$Sim'_{\text{Brushlet}}(P,Q) = \frac{1}{2} + \frac{Sim_{\text{Brushlet}}(P,Q) - \mu_{\text{Brushlet}}}{6\sigma_{\text{Brushlet}}} \quad (9)$$

$$Sim'_{LBP}(P,Q) = \frac{1}{2} + \frac{Sim_{LBP}(P,Q) - \mu_{LBP}}{6\sigma_{LBP}} \quad (10)$$

In the above equation, $\sigma_{\text{Brushlet}}$, $\mu_{\text{Brushlet}}$, $\sigma_{LBP}$ and $\mu_{LBP}$ respectively represents the standard deviation and the mean value of $Sim'_{\text{Brushlet}}$ and $Sim'_{LBP}$.

The distance between the two medical images is as follows:

$$Sim(P,Q) = w_1 Sim'_{\text{Brushlet}}(P,Q) + w_2 Sim'_{LBP}(P,Q) \quad (11)$$

In the equation, $w_1$ and $w_2$ are for the weight, and meet the formula that $w_1 + w_2 = 1$.

*E. The Algorithm of Medical Image Retrieval*

*1) The Medical Image Storage of MapReduce*

Image storage is the foundation of the automatic medical image retrieval, and it is a data-intensive computing process. The using of the traditional method to put the image into HDFS is very time-consuming, thus the distributed processing method of MapReduce is applied to upload the image to HDFS. The specific situation is as follows:

(1) In the Map phase, using the Map function to read a medical image every time, and extract the color and texture feature of the image.

(2) In the Reduce phase, the extracted feature data of medical image is stored in HDFS. HBase is a column-oriented distributed database, thus the table form of it is used for the medical image of HDFS. The specific process is shown in Figure 3.

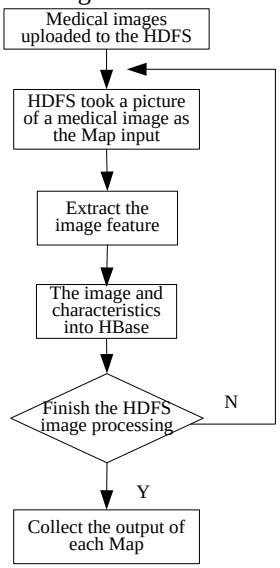


Figure 3. Storage process of medical image

Upload the medical images to HDFS→Take a medical image from HDFS and input it as Map→Extract the image features→Write the image and features in HBase →Complete the image processing in HDFS→Collect the output of Map

*2) Medical Image Retrieval of MapReduce*

The medical image and its features are all stored in HBase, when the data set of HBase is very large, the scan and search of the entire table will take a relatively long time. To reduce the time of image retrieval and improve the retrieval efficiency, the MapReduce calculation model is used to conduct the parallel computing of medical image retrieval. The specific framework is shown in Figure 4.

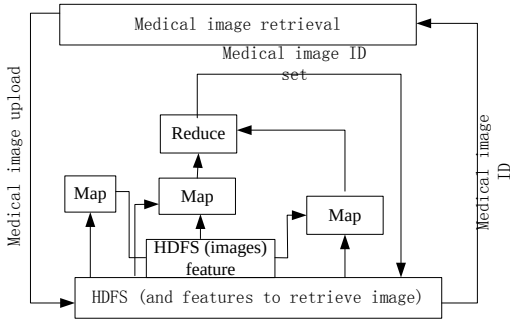


Figure 4. Work diagram of image retrieval

The steps of MapReduce based medical image retrieval are as follows:

(1) Collect the medical images, extract the corresponding features and store the features into HDFS.

(2) With the user's submission of search requests, extract the Brushlet features and LBP features of the medical images waiting for retrieval.

(3) In the Map phase, conduct the similarity matching between the features of the medical images waiting for retrieval and the features of images in HBase. The output of the map is the key value of <similarity, image ID>.

(4) Conduct the ranking and redistricting of the whole key value of <similarity, image ID > output by map, according to the size of the similarity, and then input them into the reducer.

(5) In the Reduce phase, collect all the key-value pairs of <similarity, image ID >, then conduct the similarity sorting of these key values, and write the first N keys into the HDFS.

(6) Output the ID of those images that are the most similar to the medical images waiting for retrieval, and the user gets the final result of the medical retrieval.

The function of Map and Reduce is as follows:

$Map(key, value)$

Begin

//read the features of the medical images waiting for retrieval

$Csearch = \mathrm{Re}ad\ Search Charact(\ )$;

// read the data in the feature library

$Cdatabase = value$;

// read the image path in the image library

$Path = Get\ Figure\ Path(value)$;

// calculate the similarity between the features of Brushlet domain and the features of LBP

$SimByBrushlet = Compare\ By\ Brushlet$
$(Csearch, Cdatabase)$ ;

$SimByLBP = CompareByLBP(Csearch, Cdatabase)$;

// calculate the similarity of matching, among which $w_1$ and $w_2$ respectively represents the similarity weight of the Brushlet domain features and LBP features.

$Sim = w_1 * SimByBrushlet + w_2 * SimByLBP$;

$Commit(Sim, Path)$;

End

$\mathrm{Re}duce(key, value)$

Begin

// conduct the ranking of the medical images

$Sort(key, value)$;

// key refers to the similarity value, value refers to the path of the similar medical images

$Commit(key, value)$;

End

## III. THE SIMULATION TEST

### A. Experimental Environment

Under the Linux environment, one master node (Name Node) and three work nodes (Data Node) form a Hadoop distributed system. The specific configuration is shown in table 1.In the Hadoop distributed system, by conducting the test of medical image retrieval with different number of nodes, compare its test results with the test results of the traditional image retrieval system in literature [15] and the image retrieval system under the B/S structure. The system performance evaluation criteria use the storage efficiency, retrieval speed, precision ratio (%) and recall ratio (%), and analysis the performance of the Hadoop distributed image retrieval system.

TABLE I. CONFIGURATION OF EACH NODE IN THE DISTRIBUTED SYSTEM

| Node | CPU | RAM | IP |
|---|---|---|---|
| NameNode | Intel Core i7-3770K 4.5GHz | 8G | 192.168.0.1 |
| DataNode1 | AMD Athlon II X4 631 2.8GHz | 2G | 192.168.0.21 |
| DataNode2 | AMD Athlon II X4 631 2.8GHz | 2G | 192.168.0.22 |
| DataNode3 | AMD Athlon II X4 631 2.8GHz | 2G | 192.168.0.23 |

### B. Load Performance Testing of the System

For Hadoop medical image retrieval system, the CPU usage rate of each node in 400000 medical images is shown in Figure 5. From Figure 5 it is known that due to there are only two Map tasks, the tasks are respectively assigned to DataNode1 and DataNode3. In the t1 and t2 moment, the Map tasks of the two nodes are in the execution; in t3 moment, the Map task in DataNode3 has been completed and the Reduce task is started in this node, while the Map task in DataNode1 is still in the complementation; in t4 moment, the Map task in DataNode1 is completed, and DataNode1 transfers the intermediate result generated from the Map task to DataNode3 to conduct the processing of Reduce; in t5 moment, only DataNode3 is processing the Reduce task, while DataNode1 and DataNode2 are idle; in t6 moment, the whole retrieval task is finished, each node is in the idle state. For 800000 and one million medical images, the CPU usage rate of each node is shown in Figure 6 and Figure 7. From Figure 6 and 7 it is known that the loading condition of each node is similar to that of 400000 medical images.

### C. Result of the Medical Image Retrieval

After uploading a medical image, and using the Hadoop medical image system to retrieve, the results are shown in Figure 8. From Figure 8 it is known that the retrieval results are relatively better. The results show that the Hadoop distributed medical image system is based on Hadoop, and uses the Map/Reduce method to decompose the tasks, which transforms the traditional single-node working mode into the teamwork between all the nodes in the cluster, and splits the parallel tasks to the spare nodes for processing, improves the retrieval efficiency of the medical image.

### D. Performance Comparison with the Traditional Method

#### 1) Contrast of Storage Performance

With different number of medical images, and under the situation of different nodes, the storage time of the images is shown in Figure 9. From Figure 9 it is known that, when the number of the medical image is less than 200000, the difference of the storage performance between the two systems is little. But with the increasing

of the image number, the storage time of the B/S single-node system increases sharply, while that of the Hadoop distributed system grows slowly. At the mean time, the storage performance of the text-based system is superior to that of the traditional Hadoop image processing system. This is because the traditional Hadoop image processing system is still using the traditional uploading method, which only uses the Map/Reduce method in the process of image retrieval, while the text-based system uploads the medical images to HDFS through the method of Map/Reduce.

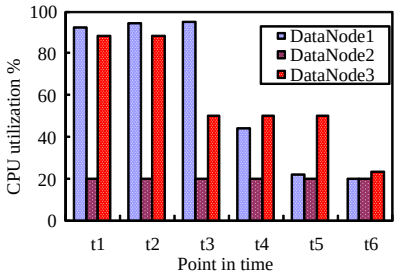


Figure 5. CPU Usage Rate of Processing 400000 Medical Images

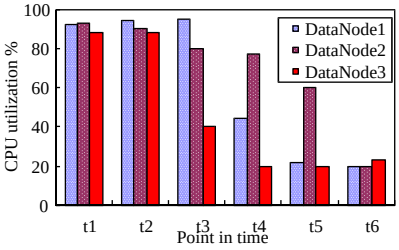


Figure 6. CPU Usage Rate of Processing 800000 Medical Images

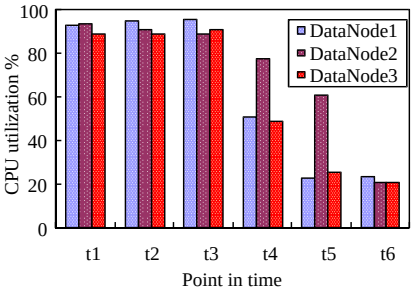


Figure 7. CPU usage rate of processing one million medical images

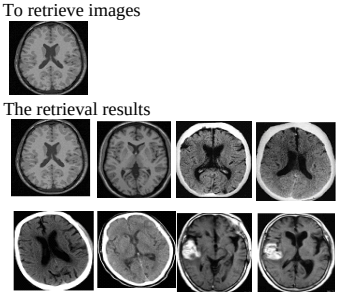


Figure 8. Result of the medical image retrieval

*2) Contrast of Retrieval Efficiency*

With different size of medical image library, under the situation of different nodes, the retrieval time of the medical images is shown in Figure 10. From Figure 10 it is known that, when the size of the medical image is small, the difference of the retrieval time between the distributed system and the B/S single-node system is little. With the increasing of the medical image's number, the retrieval time of the two systems increases accordingly. But the retrieval time of the B/S single-node system grows with larger amplitude, while that of the Hadoop medical image system grows more slowly. That is mainly due to the advantage of using the Map/Reduce parallel computing, which assigns the medical image retrieval tasks to multiple nodes, improves the retrieval efficiency of the medical images. At the same time, the more nodes there are, the faster the speed will be. By increasing the nodes of the Hadoop system, the performance of the image retrieval system is improved.

Compared with the traditional Hadoop image retrieval system, the text-based image retrieval system adopts the Map/Reduce method to conduct the parallel processing for both image storage and image matching. Relatively to the traditional Hadoop image retrieval system, which only adopts Map/Reduce method for image matching, the text-based retrieval system reduces the time to scan and search the whole medical image feature library and the time of medical image matching, improves the image retrieval efficiency.

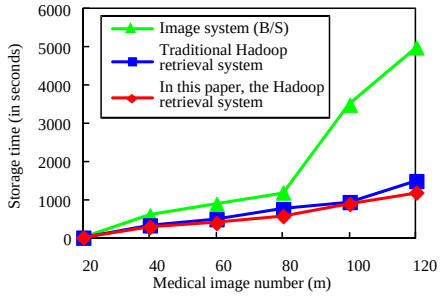


Figure 9. Storage time comparison within three systems

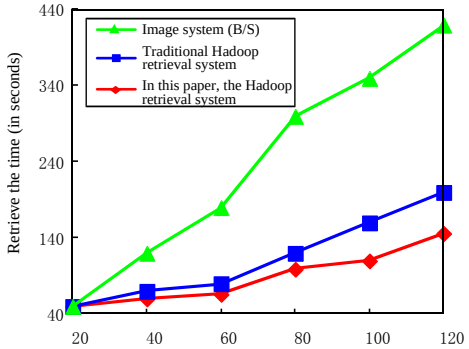


Figure 10. Medical image retrieval efficiency comparison between two systems

*3) Contrast of Retrieval Results*

For different types of medical images, by using the Hadoop and traditional retrieval system to conduct the comparison experiment, the precision rate and recall rate are shown in Table 2 and Table 3. From Table 2 and Table 3 it is known that the precision rate and recall rate of the text-based Hadoop system are slightly higher than those of the traditional Hadoop image retrieval system and B/S single-node image retrieval system, the

TABLE II. PRECISION RATE (%) COMPARISON WITHIN MULTIPLE TYPES OF MEDICAL IMAGES

| Different Types of Medical Images | Text-based Retrieval System | Traditional Retrieval System | B/S Single-node Retrieval System |
|---|---|---|---|
| Images of Brain CT | 95.04 | 94.98 | 94.63 |
| Images of Brain MRI | 91.61 | 91.58 | 91.28 |
| Images of Skin-micro | 93.67 | 92.93 | 92.26 |
| Images of X-ray Breast | 91.46 | 91.09 | 90.67 |
| HRCT of Lung | 93.52 | 92.93 | 92.53 |

TABLE III. RECALL RATE (%) COMPARISON WITHIN MULTIPLE TYPES OF MEDICAL IMAGES

| Different Types of Medical Images | Text-based Retrieval System | Traditional Retrieval System | B/S Single-node Retrieval System |
|---|---|---|---|
| Images of Brain CT | 92.21 | 91.26 | 91.59 |
| Images of Brain MRI | 90.32 | 89.84 | 90.94 |
| Images of Skin-micro | 90.38 | 90.32 | 90.33 |
| Images of X-ray Breast | 90.82 | 90.04 | 89.60 |
| HRCT of Lung | 91.10 | 90.57 | 89.31 |

advantages over the precision rate and recall rate is not obvious. But for the large-scale medical image retrieval system, the merits of the system performance are mainly measured by the image retrieval efficiency. Through Figure 10 it is known that the text-based Hadoop distributed system effectively reduces the retrieval time of the medical image, improves the retrieval efficiency of the medical image, which better solves the problem that the massive medical images retrieval has a low efficiency, obtains a relatively satisfactory retrieval results.

## IV. CONCLUSION

CBMIR medical image retrieval is a data-intensive computing process, the traditional B/S single-node retrieval system has the defects of low efficiency and poor reliability and so on. Thus, a kind of Hadoop medical image retrieval system is put forward. The results of the simulation test show that the Hadoop medical image retrieval system improves the efficiency of the image storage and image retrieval, obtains a better retrieval result, and can satisfy the real-time requirements of the medical image retrieval. Especially when deals with the massive medical images, it has the advantages the traditional B/S single-node system cannot compared with. Therefore, the working focuses in the future are improving the transmission speed of data between the Map task and the Reduce task, reducing more time consumption which is due to the transfer of information, to further improve the execution efficiency of the existing image retrieval system.

## REFERENCES

[1] Song ZHen, Yan Yongfeng. Interest points in images integrated retrieval features of the. *2012 based on computer applications*, 32 (10) pp. 2840-2842.

[2] Zhang Quan, Tai Xiaoying. Relevance feedback Bayesian in medical image retrieval based on. *Computer Engineering*, 2008, 44 (17) pp. 158-161.

[3] Yu Sheng, Xie Li, CHeng Yun. Image color and primitive features of computer application based on. 2013, 33 (6) pp. 1674-1708.

[4] FAY C, JEFFREY D, SANYJAY G, et al. Bigtable: A distributed storage system for structured data// *Proceedings of the 7th Symposium on Operating Systems Design and Implementat. Seattle: WA*, 2006, 276-290.

[5] KEKRE H B, THEPADE S, SANAS S. Improving performance of multileveled BTC based CBIR using sundry color spaces. *International Journal of Image Processing*, 2010, 4(6) pp. 620-630.

[6] Liye Da, Lin Weiwei. A Hadoop data replication method of computer engineering and applications, 2012, 48 (21) pp. 58-61.

[7] Wang Xianwei, Dai Qingyun, Jiang Wenchao, Cao Jiangzhong. Design patent image retrieval methods for MapReduce. *Mini micro system based on 2012*, 33 (3, 626-232.).

[8] SANJAY G, HOWARD G, SHUNTAK L. The Google File System// *Proceedings of the 19th ACM Symposium on Operating Systems Principles. Bolton Landing: ACM*, 2003 pp. 29-43.

[9] Liang Qiushi, Wu Yilei, Feng Lei. MapReduce micro-blog user search ranking algorithm of computer application based on. 2012, 32 (11) pp. 2989-2993.

[10] JEFFREY D, SANJAY G. Mapreduce: a flexible data processing tool. *Communications of the ACM 2010*, 53(1) pp. 72-77.

[11] KONSTANTIN S, HAIRONG K, SANJAY R, et al. Hadoop distributed file system for the Grid// *Proceedings of the Nuclear science Symposium Conference Record (NSS/MIC). IEEE: Orlando*, 2009 pp. 1056-1061.

[12] JEFFREY D, SANJAY G. Mapreduce: simplified data processing on large clusters *// Proceedings of the 6th Symposium on Operating Systems Design and Implementat. IEEE: San Francisco*, 2004 pp. 107-113.

[13] Lian Qiusheng, Li Qin, Kong Lingfu. The texture image retrieval combining statistical features of the circular symmetric contourlet and LBP. *Chinese Journal of computers*, 2007, 30 (12) pp. 2198-2204.

[14] Wang Zhongye, Yang Xiaohui, Niu Hongjuan. Brushlet domain retrieval algorithm based on complex computer simulation of. image texture characteristics, 2011, 28 (5) pp. 263-266, 282

[15] ZHANG J, LIU X L, LUO J W, BO L T N. DIRS: Distributed image retrieval system based on MapReduce// *The Network Security and Soft Computing Technologies. IEEE: Maribor*, 2010 pp. 93-98.