Remote Sensing Image Retrieval Algorithm based on MapReduce and Characteristic Information

Zhang Meng[1,2]

1 *Computer School*, Wuhan University
Hubei, Wuhan430097
2 Information Center, Wuhan University
Hubei, Wuhan430097

***Abstract* — In order to improve the retrieval efficiency and accuracy of remote sensing image, and this paper proposed a remote sensing image retrieval algorithm based on MapReduce. Firstly, the image color and texture features of emote sensing are extracted, and then the Map function is used to compute similarity among the retrieval remote sensing images and the feature library he according to color, color features, finally, the intermediate results of nodes are collected the node is obtained by using the Reduce function, and the remote sensing images are sorted to according to the similarity to obtain the remote sensing image retrieval results. Test results show that the proposed algorithm can fast and accurate retrieval in remote sensing image, not only improve the remote sensing image retrieval efficiency, and also improve the remote sensing image retrieval accuracy.**

Keywords - remote sensing image；feature extraction；cloud computing；retrieval algorithm

I. INTRODUCTION

With the development of satellite remote sensing technology, remote sensing images data increase daily, there are some disadvantages existed in the traditional manual retrieval method such as large workload and low efficiency, which could not meet requirements of remote sensing image application, while the automatic retrieval of remote sensing images based on computers could enhance retrieval efficiency and effectiveness, therefore, designing efficient and high accuracy remote sensing image retrieval algorithm has become a significant subject in the research at present.

Aiming at the automatic retrieval of remote sensing images, scholars home and abroad have conducted a large amount of researches, among which CBIR based on content has advantages of quick speed and high precision and it has become the main retrieval algorithm, firstly through drawing some characteristics of the remote sensing images such as color, type as well as texture to describe the content of the remote sensing images, then match with feature database in the remote sensing images to obtain the retrieval results [2-4]. Traditional signal node module is difficult to meet real-time requirement[5,6]. Distributed processing technology could distribute tasks to various working nodes and then treat, jointly accomplish the tasks through collaboration among nodes, therefore, distributed processing technology has provided a new kind of solution for remote sensing images retrieval[7]. Distribute processing technology at present mainly has grid computing and cloud computing, in which Hadoop is a basic architecture for distribute processing system, the user could develop MapReduce program without understanding underlying details, conducting large scale of data analysis with Hadoop has become the main parallel processing module in cloud computing and has been widely used in virtual database, large scale data processing, bio-medicine as well as classification of patent images[8].

To increase retrieval efficiency and accuracy rate of remote sensing images, this thesis has put forward a kind of retrieval algorithm of remote sensing images based on MapReduce. First of all, drawing the remote sensing images and texture features, then matching with remote sensing images according to color features with Map function, and conducting collection on intermediate results of various computing nodes with Reduce function and sorting of remote sensing images according to the similarity at last so that obtaining retrieval result of remote sensing images. The test result shows that the algorithm in this thesis could retrieve the remote sensing images fast and accurately, which not only enhances retrieval efficiency of remote sensing images but increases accuracy of retrieval of remote sensing images.

II. CHARACTERISTICS OF REMOTE SENSING IMAGE AND SIMILARITY MATCHING

Remote sensing image retrieval system based on CBIR draws remote sensing image features to be retrieved first and then compute feature similarity in remote sensing image database, realize image retrieval according to the similarity at last.

A. Drawing remote sensing image

Color is an important characteristic in distinguishing classification of remote sensing images, drawing color features of remote sensing images in RGB color space and obtaining 4 color features including RGB mean value, R mean value, G mean value as well as B mean value.

Texture describes space changes in remote sensing images and draws texture features of remote sensing images

with Gabor filter. Gabor filter h(x,y) and Fourier H(u,v) transformation forms are:

$$\begin{cases} h(x,y) = g(\hat{x}+\hat{y})\exp\left(2\pi\sigma f \hat{x}\right) \\ H(u,v) = \exp\left[\dfrac{(\hat{u}-f)^2+\hat{v}^2}{2a^2}\right] \end{cases} \quad (1)$$

$$\text{In which} \begin{cases} g(x,y) = \dfrac{1}{2\pi\sigma^2}\exp\left(-\dfrac{x^2+y^2}{2\sigma^2}\right) \\ (\hat{x},\hat{y}) = (x\cos\theta + y\sin\theta, -x\sin\theta + y\cos\theta) \\ (\hat{u},\hat{v}) = (u\cos\theta + v\sin\theta, -u\sin\theta + v\cos\theta) \\ f\times\sigma = \lambda\dfrac{(2^B+1)}{(2^B-1)} \end{cases} \quad (2)$$

$$\begin{cases} \lambda = \dfrac{\sqrt{2\ln 2}}{\pi} \\ a = \dfrac{1}{2\pi\sigma} \end{cases} \quad (3)$$

In the equation, f represents center frequency in band-pass zone of the filter, B represents tape width of the filter, θ represents direction angle of chief axis of the filter and σ represents the variance.

Determining Gabor filter parameters according to equations (1)-(3), then computing folding energy values of respective filters and images, setting the mean value and variance of image filter energy values as texture features of the remote sensing image, that is

$$F_{texture} = \{\mu_{0,0}^{texture}, \sigma_{0,0}^{texture}, \cdots, \mu_{k-1,l-1}^{texture}, \sigma_{k-1,l-1}^{texture}\} \quad (4)$$

In the equation, K represents number of center frequency, L represents number of direction angle.

Computational formula for energy mean value μ of sub image and mean square deviation σ

$$\begin{cases} \mu_{k,l}^{texture} = \dfrac{\sum_x \sum_y E_{k,l}(x,y)}{n\times n} \\ \sigma_{k,l}^{texture} = \sqrt{\dfrac{\sum_x \sum_y E_{k,l}(x,y) - \mu_{k,l}^{texture}}{n\times n}} \end{cases} \quad (5)$$

Therefore, 24 texture features of remote sensing images have been obtained and then there are 28 remote sensing image features totally composed of color and texture features.

B. Similarity matching

Suppose that the remote sensing image to be retrieved is p0, there are n images pi(i=1,2,…,n) in the remote sensing image database, its color feature is shown as ci□Rm and texture feature ti□Rk, M and K are dimensions for color and texture respectively, computing similarity between p0 and pi(i=1,2,…,n) according to the formula (6)

$$R_{0i} = w_1 D_{t_i} + w_2 D_{c_i} \quad (6)$$

In the equation, w_1 and w_2 are weights and $w_1 + w_2 = 1$, D_{t_i} and D_{c_i} show the similarity values between the color and the texture respectively, their computational formula is as follows:

$$\begin{cases} D_{t_i} = 1 - \dfrac{\left(\sum_{m=1}^{M}(t_0^m - t_i^m)^2\right)^{1/2}}{\max\limits_i \left(\sum_{m=1}^{M}(t_0^m - t_i^m)^2\right)^{1/2}} \\ D_{c_i} = 1 - \dfrac{\left(\sum_{k=1}^{K}(t_0^k - t_i^k)^2\right)^{1/2}}{\max\limits_i \left(\sum_{k=1}^{K}(c_0^k - c_i^k)^2\right)^{1/2}} \end{cases} \quad (7)$$

Conducting sorting on images in the remote sensing image database on Roi(i=1,2,…,n) in descending order and selecting previous m image as the retrieval result.

III MAPREDUCE REMOTE SENSING IMAGE RETRIEVAL

A. MapReduce image storage

Image storage is the foundation for the automatic retrieval of remote sensing image, it is a computing process in data intensive type, this thesis adopts MapReduce distributed processing to upload images to HDFS. The specific content is as follows:

(1) Map stage. Adopting Map function and reading one remote sensing image each time and then drawing color and texture features of image.

(2) Reduce stage. Storing feature data of remote sensing image drawn into HDFS. HBase is a contributed database facing rows, therefore, HDFS remote sensing image storage adopts HBase table format, specific design of HBase table is shown in table 1.

TABLE 1.HBASE TABLE DESIGNING OF REMOTE SENSING IMAGE

Remote sensing image id	Original document of image	Color feature	Texture feature
001	file001	c1	t1
002	file001	c2	t2
…	…	…	…
00n	file00n	cn	tn

Procedure for image storage based on MapReduce is shown in figure 2.

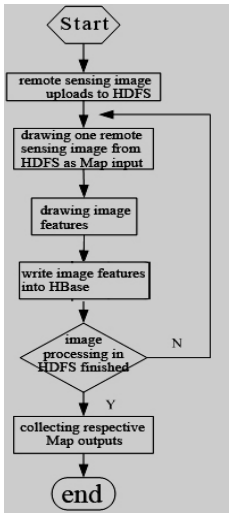

Figure 2.Storage procedure for remote sensing image

B. MapReduce remote sensing image retrieval

Because the remote sensing image and its features are stored in HBase, when HBase data collection is so large, long time shall be spent on scanning the table as a whole. To reduce time for image retrieval and enhance retrieval efficiency, conducting parallel computing on remote sensing image retrieval with MapReduce computing module, the specific frame is shown in figure 3 and specific implementation process is shown in figure 4.

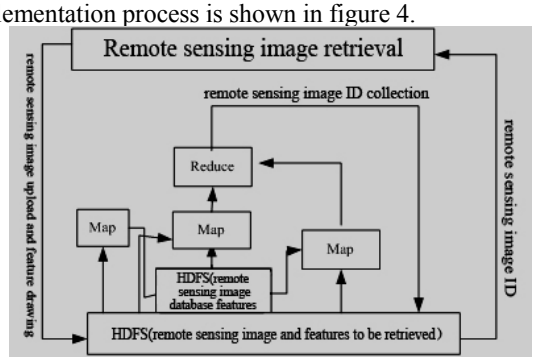

Figure 3.Working procedure for remote sensing image retrieval

Steps for remote sensing image retrieval based on MapReduce are as follows:

Step 1: Map stage. Read remote sensing image to be retrieved from HDFS cache and draw its color and texture features, then match with features in image in HBase, map output is the value of <similarity, image ID >.

Step 2: Conducting sort and redraw of all values of map outputs <similarity, image ID > and then input to reducer again.

Step 3: Reduce stage. Collecting all of values of <similarity, image ID > and then conducting sort of similarity on these values and writing N values into HDFS.

Step 5: Outputting those image IDs that are the most similar to the remote sensing images to be retrieved.

Map function is defined as:

```
map(key,value)
Begin
Csearch=ReadSearchCharact( ); //read features of remote sensing image to be retrieved
Cdatabase=value; //read data in remote sensing features database
Path = GetPicturePath( value) ; / /read image route in remote sensing image database
SimByColor=CompareByColor(Csearch, Cdatabase) ; / /computing similarity of remote sensing image color
SimByTexture = CompareByTexture(Csearch, Cdatabase); //computing similarity of remote sensing image texture
Sim=SimByColor*w1 + SimByTexture*w2; //computing matching similarity
Commit(Sim,Path);
End
```

Reduce function is defined as

```
reduce(key,value):
Begin
Sort(key,value); //conducting sort on remote sensing image according to size of similarity
Commit(key,value); //key refers to the value of similarity ， value refers to route of similar remote sensing images
End
```

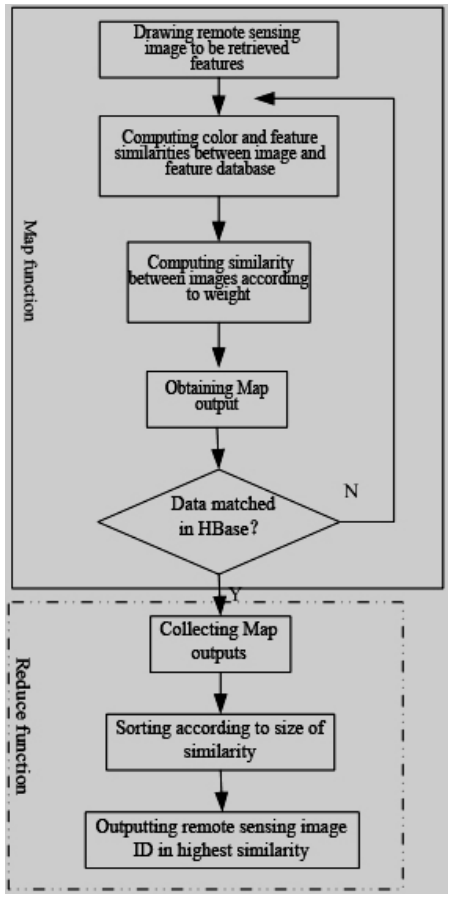

Figure 4.Process of remote sensing image retrieval based on MapReduce

IV. SYSTEM TEST AND ANALYSIS

A. Test environment

Adopting one main engine and 3 ordinary machines to consist of one Hadoop distributed system through Linux environment and their configuration is shown in table 2. There are 2000 remote sensing images collected totally. To make the result of remote sensing image retrieval put forward in this thesis more convincing, we conducted contrast experiment adopting B/S single node system.

TABLE 2. CONFIGURATION OF VARIOUS NODES

Nodes	Operation system	IP	CPU	RAM
Main engine	Linux	192.168.0.101	Core i7 3960X 3.3GHz	4G
Ordinary1	Linux	192.168.0.102	Core i3 2120 3.3GHz	2G
Ordinary2	Linux	192.168.0.103	Core i3 2120 3.3GHz	2G
Ordinary3	Linux	192.168.0.104	Core i3 2120 3.3GHz	2G

B. Test analysis on storage performance

Adopting different amount of remote sensing images and the storage time of images under different nodes is shown in figure 5. It can be seen from figure 5 that when the amount of remote sensing images is less than 500, there is no big difference in storage time between B/S single node system and Hadoop distributed system and the advantage is not obvious. When the amount of remote sensing images is more than 500, storage time in B/S single node system has increased greatly while slow in Hadoop distributed system, this shows that uploading remote sensing images into HDFS with MapReduce method will enhance storage efficiency. When the amount of images is more than 2000, storage time in 2 nodes and 3 nodes distributed system show increase in index form, this has shown that Map tasks are more than 3 at this time meanwhile it will distribute many tasks on some nodes, however, one node can only execute one Map task in one time, so it increases number of nodes in Hadoop distributed system which enhance execution efficiency of remote sensing image retrieval system.

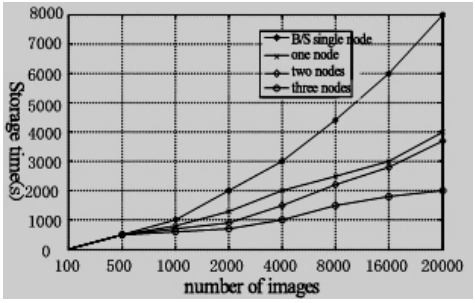

Figure 5. Change curve of storage time for remote sensing image.

C. Test analysis on remote sensing image retrieval

Remote sensing image retrieval time consumption in different scale of remote sensing image database under different nodes is shown in figure 6. It can be seen from figure 6 that when the amount of images in remote sensing image database is small, multi-node retrieval time in Hadoop distributed system is longer than that in B/S single node system and one node system, it is mainly because conducting parallel computing adopting multi-node and increase in the amount of calculation and time, when the number of images is more than 1000, retrieval time of images in multi-node distributed system is obviously less than the single node, it is mainly because advantage in conducting parallel computing with MapReduce to distribute the task of remote sensing image retrieval to various nodes which increases efficiency of remote sensing image retrieval.

D. System load test

Under 3 nodes, forwarding remote sensing image retrieval task to Hadoop distributed system, testing load conditions of various nodes under different time points and different amounts, recording CPU utilization ratios of various nodes are shown in figures 7 and 8 respectively.

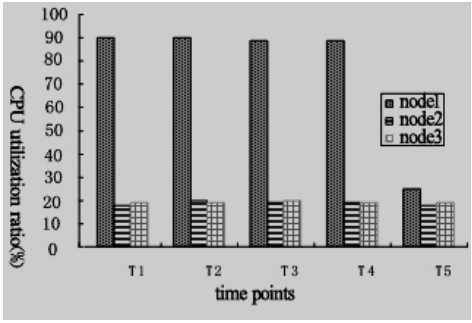

Figure 7. CPU utilization ratio in processing 200 remote sensing images

It can be seen from figure 7 that when the amount of images processing(200)due to small amount of images and only one Map task, it distributes to node1 to process and finishes at t5, node1 begins to execute Reduce task.

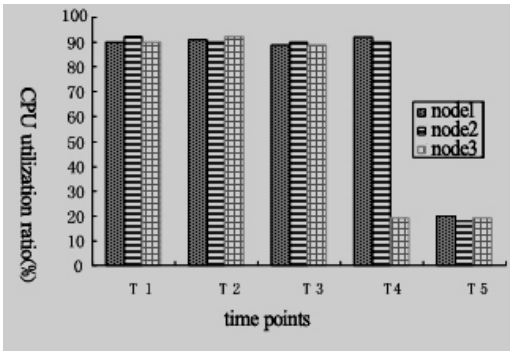

Figure 8.CPU utilization ratio in processing 2000 remote sensing images

It can be seen from figure 8 that when the amount of remote sensing images processing is large(2000), because there are many Map tasks to be executed at the same time, 3 Map tasks on 3 nodes at T1 and T3 have been finished, because Map task on node3 at T4 has been finished and is free, therefore executing Reduce task on node3 and this has realized execution of automatic transfer of node task in

heavy load to free node, which has kept balance of system loads. Meanwhile due to collaboration between Map and Reduce tasks, it has full taken advantage of data processing capacity on various nodes and enhanced data efficiency of various nodes.

E. Comparison of results of remote sensing image retrieval

Conducting retrieval on many categories of remote sensing images with Hadoop distributed system and B/S single node system, the average retrieval results are shown in table 3. It can be seen from the table that precision ratio and recall ratio of Hadoop distributed system are superior to B/S single node system, which shows that Hadoop distributed system has enhanced retrieval quality of remote sensing images.

TABLE 3.COMPARISON OF RESULTS OF MANY CATEGORIES OF REMOTE SENSING IMAGES

Different categories	Hadoop distributed system		B/S single node system	
	precision ratio(%)	recall ratio(%)	precision ratio(%)	recall ratio(%)
Plantation	93.36	77.96	91.50	76.92
Wasteland	87.61	79.99	86.44	77.18
Houses	81.96	70.89	79.52	69.30
Lakes	84.37	67.86	82.33	66.59
Rivers	75.80	65.31	74.97	64.24
Roads and squares	81.05	60.53	79.41	58.74

VI. CONCLUSION

Aiming at enormous amount of difficulties of remote sensing image retrieval efficiency in traditional methods, this thesis has put forward a remote sensing image retrieval algorithm based on MapReduce with the advantage of Hadoop distributed technology. The test result shows that the algorithm in this thesis could retrieve remote sensing images fast and accurately, which not only enhances retrieval efficiency of remote sensing images but increases retrieval accuracy of remote sensing image and has wide application prospect in automatic retrieval of remote sensing images.

REFERENCES

[1] Li Chao Feng, Zeng Sheng Gen, Xu Lei, intelligent processing of remote sensing image , Beijing: Electronics Industry Press, 2007,pp.99-103.

[2] Simpson, J. J., J. T. Mcintir. A Recurrent Neural Network Classifier for Improved Retrievals of Area Extent of Snow Cover. IEEE Transactions on Geosciences and Remote Sensing, 2001, 39,pp. 2135-2147.

[3] Smeulders A W.M., Worring M, Santini S, et al. Content -based image retrieval at the end of the early years. IEEE Trans. On Pattern Analysis and Machine Intelligence. 2000, 22(12),pp.1-32.

[4] Guo Zhi Qiang Cai Song Classification algorithm of colorful remote sensing image and Matlab realization Wuhan Science and Engineering University learned journal, 2006,28(1), pp. 108-111.

[5] Wang Xian Wei, Dai Qing Yun, Jiang Wen Chao, Cao Jiang Zhong. Retrieval method for appearance design patent image.Mini-size computer system, 2012, 33(3), pp.626-232.

[6] Sanjay Ghemawat,Howard Gobioff,Shun-Tak Leung.The Googl File System. Pro-C eedings of the 19th ACM Symposium on Operating Systems Principles.Bolton Landing:ACM,2003,12,pp.29-43.

[7] Jeffrey Dean , Sanjay Ghemawat. MapReduce: a flexible data processing tool. Communications of The Acm, 2010, 53(1), pp.72-77.

[8] Tian Xia. Large-scale SMS messages mining based on MapReduce. Proceedings of the International Symposium on Computational Intelligence and Design, London, 2008,13,pp. 7-12.

[9] Konstantin Shvacliko, Hairong Kuang, Sanjay Radia, et al. Hadoop distributed file system for the Grid. Proceedings of the Nuclear science Symposium Conference Record, IEEE, 2009,pp.1056-1061.