International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 7 Issue 11 Ver I || Nov 2018 || PP 73-77

Image Deduplicationsystem Based On Cloud Computing

Qin Wang[1], Qianjin Wang[2],Jun Ye[3],Xudong Lin[3],Liufen Li[3]*
[1](School of Automation & Information Engineering, Sichuan University of Science & Engineering, Sichuan, China)
[2](School of Biotechnology, Sichuan University of Science & Engineering, Sichuan, China)
(School of Mathematic and Statistics, Sichuan University of Science & Engineering, Sichuan, China)
Corresponding author;Liufen Li

Abstract :*With the cloud computing, people paid more and more attention to cloud storage. The main problems are insufficient storage space, large network transmission load and so on. In order to solve the above problems, this paper proposes a cloud-based image retrieval verification system, and image similarity processing in the cloud, not only can increase the user's use and experience, but also reduce the load of the network transmission and reduce the data redundancy of the server. Under the premise of guaranteeing the security of network information, not only ensure the user's privacy and the use of comfort, but also ensure that customers under the cost of many users. This paper not only has the very high theory research value, also has the high promotion use value in the practical application.*

Keywords *-cloud computing,imageDeduplication,storage*

Date of Submission: 12-11-2018 Date of acceptance: 26-11-2018

I. INTRODUCTION

With the progress of the Times the network has deepened our lives, already and our life is inseparable, for our life has brought great convenience, more and more data into the cloud.The capacityeach user is limited, however, the capacity of cloud computing is powerful [1-5]. The data is getting more and more, and data redundancy is come out.So in the face of such a large environment, saving cloud capacity is imminent.

Data deduplication[6-9] in cloud storage environment can effectively reduce the communication overhead between cloud storage users and cloud servers, and reduce the storage overhead of cloud storage servers[10-13]. Therefore, from an economic point of view, cloud storage can be re optimized to achieve optimal resource allocation in cloud storage.

II. PRELIMINARIES

2.1 LSH Basic Ideas

In real life, similar data is to be judged, mostly in high-dimensional data, if the use of linear lookup is not realistic, so LSH is to reduce the data dimension, and to ensure that the original dimension of similar data, dimensionality reduction after similar probability is also very large, but not similar to the data is very small.

2.2Minimum hash

The minimum hash is defined as: The row number of the row with the first column value of 1 after the feature matrix is randomly arranged in rows.

The probability that the two columns ' minimum hashes are equal is the same probability as the jaccard of two columns.

Demonstrate:

Divides the possible results of the data into three categories:

(1) Class A: The values for both columns are 1;
(2) Class B: One of the columns has a value of 0, and the other column has a value of 1;
(3) Class C: Both columns have a value of 0.

Feature matrices are fairly sparse, causing most rows to belong to Class C, But only Class A and Class B can determine the sim (si,sj), assuming A-class has a, B-class row has b, then the sim (si,sj) =a/(a+b). Now all we need to do is show the probability that the minimum hash value of two is equal when the row of the matrix is randomly arranged.

P(h(Si)=h(Sj))=a/(a+b)

If we delete the Class C row, then the first row is a Class A row or a class B row, if the first row is class A, then h (Si) =h (Sj), so P(h(Si)=h(Sj))=P(After the class C row is deleted, the first line is Class A)= Number of Class A rows/Number of all rows =a/(a+b). Multiple min-hashing, the signature matrix is obtained, and the similarity of the minimum signatures of the two columns is the estimation of the Jaccard similarity of the two columns.

III. PRINCIPLES

3.1 Implementation mode

3.1.1 Index construction

(1) The similarity is calculated by using Jacquard distance.

(2) Select LSH hash functions to satisfy (D1, D2, P1, P2)-sensitive; the signature matrix is obtained by using the minimum hash (min-hashing) multiple times.

(3) According to the accuracy of the search results (that is, the probability that the adjacent data is found), the signature matrix is divided into B group and R row data in each group.

(4) All groups are built into their own bucket space (total B), and each group is individually minhashing, and the corresponding bucket number is obtained.

3.1.2 Divide the bucket number

According to the signature matrix, divided into Group B, each group is divided into R rows, then each group is min-hashing to get the bucket number of each group. The values of B and r determine the similarity of the jaccard between the two signatures and the two columns. Including:

(1)Two signature values for all rows in a group are equal probability is p^r.

(2)The probability that at least one pair of signatures in a group is not equal is $1 - p^r$.

(3)The probability that at least one pair of signatures in each group is unequal is $(1 - p^r)^b$.

(4)The probability of having at least one group of all pairs of signatures equal is$1 - (1 - p^r)^b$.

3.1.3. Search

(1) The corresponding bucket number is obtained by LSH hash function hash of the query data;

(2) Take out the corresponding data in the bucket number, (in order to ensure the search speed, usually only need to take out the first 2 L data);

(3) Calculates the similarity or distance between the query data and the 2 L data, and returns the nearest neighbor data.

3.1.4 Deduplicaiton

In the online query phase, if the data to be queried and one of the data, in B bucket space, all the barrel numbers are the same, it is judged that there is a possibility for the same picture, and the corresponding picture in the database, random interception of ciphertext with cosine similarity for interpretation, if the same is interpreted as the same picture, do not store. The process of deduplicaiton is shown in Fig.1.

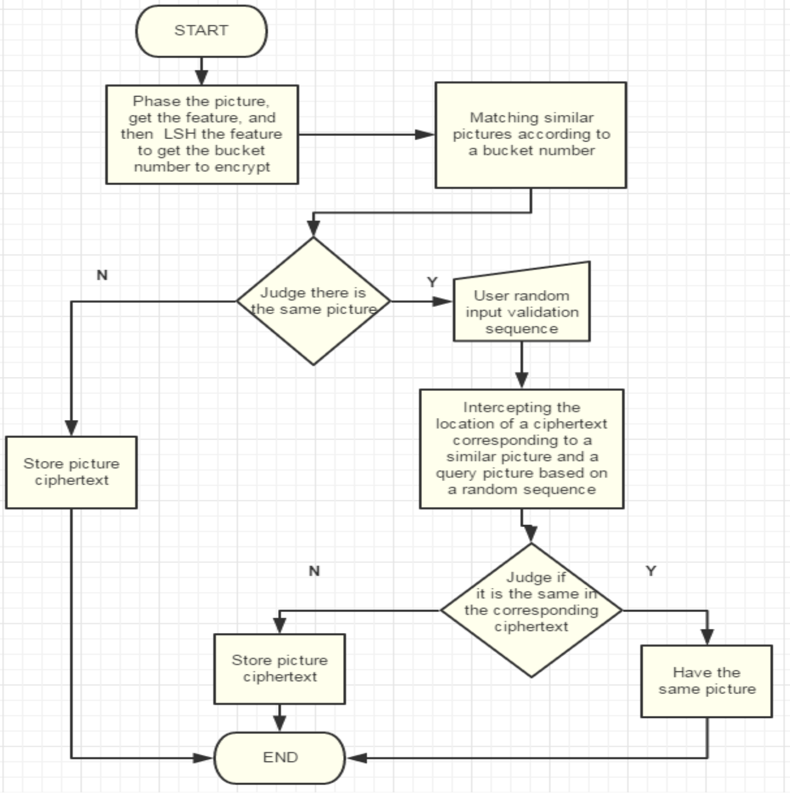

START
Phase the picture, get the feature, and then LSH the feature to get the bucket number to encrypt
Matching similar pictures according to a bucket number
N
Judge there is the same picture
Y
User random input validation sequence
Intercepting the location of a ciphertext corresponding to a similar picture and a query picture based on a random sequence
Store picture ciphertext
N
Judge if it is the same in the corresponding ciphertext
Y
Store picture ciphertext
Have the same picture
END

Fig.1 Deduplication

IV. IMPLEMENT

4.1 Test environment

Client: Windows7, intelcorei5-4210M@2.6GHz, 8G Memory , ADATASP900128GB, java1.8
Server: CentOS6.564CPU, 1 core, 2G RAM,Bandwidth 1M, System disk 40G, Java1.8. (Aliyun Server)

4.2 Test environment

(1) Step: To install the Java JDK: http://www.oracle.com/technetwork/java/javase/downloads/index.html, Log on to the Web site to download the latest JDK, recommended 1.8 versions.

(2) Then double-click to open, enter the default open, has been the default click Next, until completed. Then the JDK installation was successful.

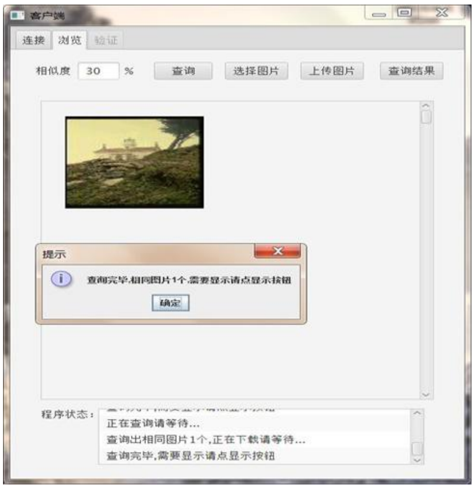
(3) Configure environment variables, right-click My Computer, select Properties-> Advanced system Settings-> Select System environment variable-> New system environment variable(The variable value is the Lib folder path under the Java installation path).

(4) In the system environment variable, click "Path" editor, add “%JAVA_HOME%/bin;” to the front, click OK to complete.

(5) At this point, the JDK installation completed, we have to verify that the installation is correct, win+R or click Start--〉 run Input "cmd", open the System command Prompt box, input "java-version", appear as shown in the picture, indicating the installation was successful.

4.3 Testing

The system query picture function is to use the bucket number of picture to the server's picture library to retrieve, achieves the fast retrieval the goal, and before executes the query must set the picture the similarity degree. The interface after the search succeeds is shown in Figure 2:



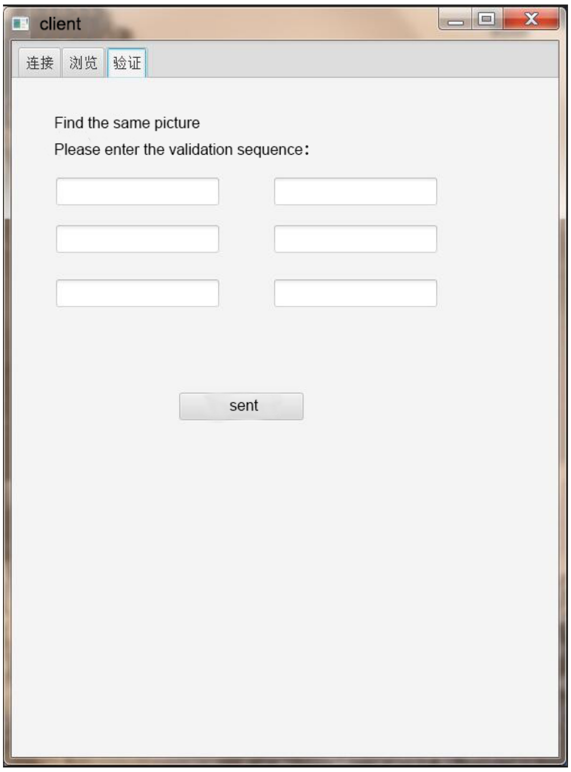
相似度 30 %

查询完毕,相同图片1个,需要显示请点显示按钮

程序状态：
正在查询请等待...
查询出相同图片1个,正在下载请等待...
查询完毕,需要显示请点显示按钮

Figure 2 Search

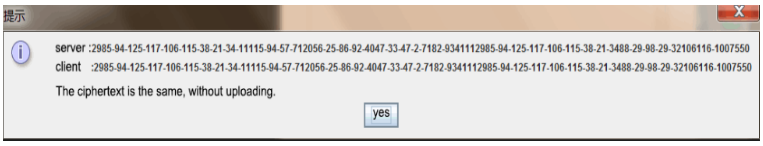
When uploading pictures, the system will automatically weight the pictures, when the server thinks it is possible to have the same picture, it prompts the user to verify the same picture through the input verification sequence. The verification interface is shown in Figure 3 and Figure 4.



Find the same picture

Please enter the validation sequence：

Figure 3 Input validation Sequence



server :2985-94-125-117-106-115-38-21-34-11115-94-57-712056-25-86-92-4047-33-47-2-7182-9341112985-94-125-117-106-115-38-21-3488-29-98-29-32106116-1007550

client :2985-94-125-117-106-115-38-21-34-11115-94-57-712056-25-86-92-4047-33-47-2-7182-9341112985-94-125-117-106-115-38-21-3488-29-98-29-32106116-1007550

The ciphertext is the same, without uploading.

Figure 4 Validation results

V. CONCLUSION

In this paper, the verification of the image is removed ciphertext, and then use the cosine distance to determine again, in order to achieve the image to weight; finally in the image retrieval module is to retrieve the data through LSH hash to achieve the efficient retrieval of the image. The main implementation of the function is through the cloud storage for individuals or enterprises to provide image upload and retrieval services, which is to increase the image of the heavy and cloud storage requirements. In short, the system realizes the user in the client to encrypt the picture, in the server side carries on the secret text storage, simultaneously has reached the protection user picture confidentiality and the data needs to be heavy.

Acknowledgements

This work was supported by the college students innovation and entrepreneurship training project (201710622037); the Talent Project of Sichuan University of Science & Engineering (2017RCL23); the Science Founding of Artificial Intelligence Key Laboratory of Sichuan Province (2017RZJ03); the Opening Project of Sichuan Province University Key Laboratory of Bridge Non-destruction Detecting and Engineering Computing (2017QYJ03);Scientific research project of Sichuan education department (14ZB0221).

REFERENCES

[1]. Li, J., Zhang, Y., Chen, X., & Xiang, Y. Secure attribute-based data sharing for resource-limited users in cloud computing.Computers & Security,72, 2018, 1-12.

[2]. Stergiou, C., Psannis, K. E., Kim, B. G., & Gupta, B. Secure integration of IoT and cloud computing.Future Generation Computer Systems,78, 2018, 964-975.

[3]. Jiang, Q., Ma, J., & Wei, F. On the security of a privacy-aware authentication scheme for distributed mobile cloud computing services.IEEE Systems Journal,12(2), 2018, 2039-2042.

[4]. Varatharajan, R., Manogaran, G., & Priyan, M. K. A big data classification approach using LDA with an enhanced SVM method for ECG signals in cloud computing.Multimedia Tools and Applications,77(8), 2018, 10195-10215.

[5]. Li, J., Zhang, Y., Chen, X., & Xiang, Y. Secure attribute-based data sharing for resource-limited users in cloud computing.Computers & Security,72, 2018, 1-12.

[6]. Sookhak, M., Gani, A., Khan, M. K., & Buyya, R. Dynamic remote data auditing for securing big data storage in cloud computing.Information Sciences,380, 2017, 101-116.

[7]. Chen, Q., Wan, Y., Zhang, X., Lei, Y., Zobel, J., & Verspoor, K. Comparative Analysis of Sequence Clustering Methods for Deduplication of Biological Databases.Journal of Data and Information Quality (JDIQ),9(3), 2018, 17.

[8]. Liu, J., Wang, J., Tao, X., & Shen, J. Secure similarity-based cloud data deduplication in Ubiquitous city.Pervasive and Mobile Computing,41, 2017, 231-242.

[9]. Fu, Y., Xiao, N., Jiang, H., Hu, G., & Chen, W. Application-Aware Big Data Deduplication in Cloud Environment.IEEE Transactions on Cloud Computing, (1), 2017, 1-11.

[10]. Yu, Y., Au, M. H., Ateniese, G., Huang, X., Susilo, W., Dai, Y., & Min, G. Identity-based remote data integrity checking with perfect data privacy preserving for cloud storage.IEEE Transactions on Information Forensics and Security,12(4), 2017, 767-778.

[11]. Yang, J., He, S., Lin, Y., & Lv, Z. Multimedia cloud transmission and storage system based on internet of things. Multimedia Tools and Applications,76(17), 2017, 17735-17750.

[12]. Li, Y., Gai, K., Qiu, L., Qiu, M., & Zhao, H. Intelligent cryptography approach for secure distributed big data storage in cloud computing.Information Sciences,387, 2017, 103-115.

[13]. Li, J., Lin, X., Zhang, Y., & Han, J. KSF-OABE: outsourced attribute-based encryption with keyword search function for cloud storage.IEEE Transactions on Services Computing,10(5), 2018, 715-725.

Liufen Li"Image Deduplicationsystem Based On Cloud Computing "International Journal of Engineering Science Invention (IJESI), vol. 07, no. 11, 2018, pp 73-77