International Journal of Engineering Research
Volume No.7, Issue No.12, pp : 199-202

ISSN:2319-6890 (online),2347-5013(print)
1 Dec. 2018

A Deduplication Scheme of Ciphertext Images

Mengxue Li, Liufen Li*

School of Mathematic and Statistics, Sichuan University of Science & Engineering, Sichuan, China

Corresponding author: Liufen Li, liliufen@suse.edu.cn

Abstract：In this paper, the bandwidth consumed by users' uploaded files is saved by decreasing in mass before uploading files. In this paper, the deduplication technique is introduced. Firstly, the plain-text keywords are extracted from the uploaded plain-text images and the index is established. And the private key of image owner and public key of image sharer are used to generate keywords and index in cipher-text index respectively. Then the public key of the image sharer is used to encrypt the plain-text image, and random numbers and cipher-text images are extracted from them. Finally, the encrypted random number, message digest and the cipher-text index of private key encryption are outsourced and stored in the personal cloud server. When the image is transmitted to the sharer again, the message digest obtained by the same random number is compared to determine whether the message digest is the same to achieve the purpose of delete-repetition.

Keyword: keyword search, EIGmamal public key system, Hash function

1. Introduction

Since 2007, the world's largest search engine service providers to Google for the first time put forward the concept of cloud computing, Cloud computing[1-4] has realized people's long-term dream of taking computing as an infrastructure, and represents the trend of information field's rapid development towards intensive scale and professional road, which has become the focus of common concern of industry, academia and government. As an extension and development of cloud computing concept, the core is the safe storage and management of data, and providing users with a certain type of storage service and access service. However, as cloud storage makes data independent of the physical control of data owners, the security, reliability and availability of cloud storage services face great challenges, and a large amount of data is also stored repeatedly, which greatly wastes the storage resources of cloud servers. Data deduplication technology[5-9] is a very important method of data management and storage optimization in cloud storage: data delete-repetition technology can eliminate data redundancy and keep only one physical copy of the same file, so as to effectively reduce the bandwidth of user-side data upload and save storage space on the server side[10-12].

In the traditional encryption system, users use different keys to encrypt the data, so that even the same data will get a completely different cipher text, so data delete-repetition cannot be implemented. Therefore, how to design a data security delete-repetition technology in an efficient cipher text environment is of great significance. As a representation of user data, more than 500 million images are uploaded every day around the world, with a large number of images being stored repeatedly. Therefore, this paper takes image as the object to design a security delete-repetition scheme of cipher text image.

2. System Model

The deduplication system model is as follows.

(1)The owner of the image uses his own plain text image to extract the plain text keyword set and establish the plain text index. Then the public key and the clear key set of the image sharer are used to generate the key a and cipher text index a .After encrypting the plain-text image using the public key of the image sharer, it is stored on the cloud server with the cipher text index a outsourcing (searching the image service for the image sharer).

(2) After the image owner uses his own plain-text image to extract the plain-text keyword set and establish the index, can simultaneously use his private key and plain-text keyword set to generate the cipher-text keyword b and index b .Then the selected random key is encrypted with its private key, and is outsourced and stored on the personal cloud server together with the cipher text index b and the message digest generated by the Hash function.

(3) The owner image transmission image x to image sharing again, first of all, from the personal cloud server get random key k and the message digest h : image owners use to generate the trapdoor information of key generation trapdoor information, and will the trapdoor information sent to the personal cloud server, cloud server receives the trapdoor information for relevant search operation (this process is not data decryption operation), and searching for related cipher image data is sent to the owner, owners use the decryption key to decrypt the document image, and is obtained k and h .

(4)The cipher text obtained after the image x is encrypted with public key (p,g,b) and random key k is processed by Hash function, and the message digest h_1 can be obtained. If $h=h_1$, the cloud server has the same image, the image x is not uploaded; Otherwise upload the image x to the image sharer.

The main process is shown in Figure 1.

ISSN:2319-6890 (online),2347-5013(print)
1 Dec. 2018

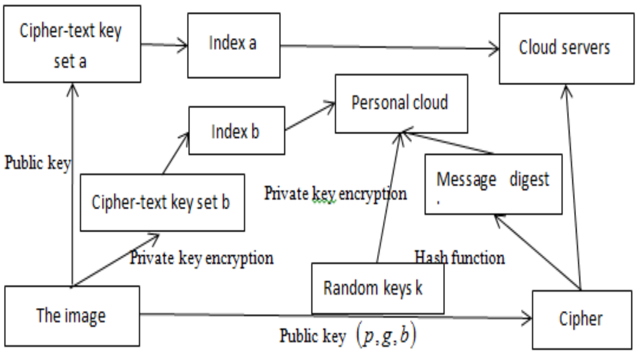

Fig.1 System model

3. The Proposed Scheme

In the process of ciphertext image delete-repetition, public key searchable encryption scheme, verifiable fuzzy keyword search scheme and SHA-3 standard Hash function are needed.

3.1 Public key searchable encryption scheme:

In this scheme, the image sharer's secret key is constructed using the ElGamal public key system. The security of the ElGamal encryption algorithm in the ElGamal public key system is based on the difficulty in solving the discrete logarithm problem.To find the private key for a given public key a , an attacker must be able to solve this discrete logarithm problem.

(1) Discrete logarithm problem on finite domain F_P : given a prime number p and a primitive element g on F_P , for $y \in F_P^*$, seeking the integer x , $0 \le x \le p-2$, make $y = g^x \bmod p$ true.

For $y = g^x \bmod p$, a given sum g , x and p , it's easy to calculate y ;But if you know the sum y , g and p , when it's a large prime number for p , it's difficult to find one x that makes the equation $y = g^x \bmod p$ true.

The key of ElGamal public key system is generated by selecting a large prime number p .Then calculate the primitive element g for Z_P^* ,selecting a smaller random number a $(a < p-1)$;When you have these three Numbers, you calculate $b = g^a (\bmod p)$;The public key consists of (p,g,b).

(2) Generator number (primitive element)

Generator numbers g are associated with prime numbers p . The way to find prime numbers p and their primitive elements g is to construct the known factorization form for $p-1$.

The steps to implement this approach are as follows:

1)Select a random number n , which will determine the number of prime factors for $p-1$;

2) Select n random prime number $q_1, q_2, ..., q_n$;

3) Select $n+1$ random number e_0 , e_1 , e_2 ,..., e_n as an index;

4) Define prime number p as $p = \left(2^{e_0} q_1^{e_1} q_2^{e_2} \cdots q_n^{e_n}\right) + 1$;

5) Verify p whether it is a prime number.

If it's prime, there's a decomposition factor $p-1$ that gives you the generator number.

(3) Encryption and decryption algorithm

If the image owner wants to send a message to the image sharer, first finding out the image sharer's public key (p,g,b) and representing the image with an integer number m $(0 \le m \le p-1)$, then choosing a random key. With these Numbers, image owners can calculate the following two Numbers:

$$C_1 = g^k$$

$$C_2 = mb^k$$

At the same time, the calculated ciphertext $C = (C_1, C_2)$ is sent to the cloud server; When the image sharer searches for the image to receive the ciphertext, the private key a can be used to restore the plain text through the following calculation.

$$C_2 C_1^{-a} = mb^k \left(g^k\right)^{-a} = mg^{ak} g^{-ak} = m \bmod p$$

$$m = C_2 C_1^{-a} \bmod p$$

3.2. Verifiable fuzzy keyword search scheme

(1) Keygen: the algorithm is performed by the image owner to generate the index and generate the key sk and private key k . Keygen algorithm is a random key generation algorithm, and the key is generated in such a way: $sk, k \xleftarrow{R} \{0,1\}^{\lambda}$,where: λ is the security parameter;

(2) Index generation: this algorithm is performed by image owners. In order to improve the search efficiency, a multi-fork tree is adopted to build the index. The core idea is that all trap doors that share a prefix have a common node, the root node is associated with an empty set, and the symbol in a trap door can be recovered from the search from the root node to the trap leaf node. All fuzzy keywords can be found through a deeply optimized search.

Suppose $\Delta = \{\alpha_i\}$ is a predefined set of symbols, the number of different symbols is $|\Delta| = 2^n$;

1) Initialization:

Image owners scan images and build keyword sets W ;

The image owner submits the encrypted image collection C to the cloud server, and the server returns the stored address ID $\{F_i\}$ of each document.

2) Fuzzy keyword set:

In order to establish an efficient set of fuzzy keywords, we use wildcard technology to construct fuzzy keywords.

3) Create the symbol - based index tree

The image owner computes the trap door $T_{w_i'} = f\left(sk, w_i'\right)$ for all fuzzy keywords, and then divides each trap door into l/n parts, each part represented by a letter in Δ; Where: l is the output length of hash function $f(x)$.

The image owner creates an index tree G_W containing all fuzzy keywords, each node contains a binary group (r_0, r_1), r_0 is the symbol value of the storing node, and r_1 stores a globally unique value $path\|mem\|F_k(path\|mem)$;Where: $path$ represents the sequence of symbols on the path from the root node to the current node, mem is a bit string that length is 2^n, representing the child node information of the current node.

Image owners attach $\left\{ID_{w_i} \middle| gk\left(ID_{w_i}\right)\right\}_{1\le i\le p}$ to the index tree G_W and outsource to the cloud server.

(1) Trapdoor $(sk, S_{w,d})$: supposing an authorized user input w is used as search input, and the default edit distance is k ;

1) The owner of the rest image adopts the corresponding algorithm to generate the fuzzy keyword set $S_{w,k}$.

2)The owner of the image computes the trap door $T_{w'} = f(sk, w')$($w' \in S_{w,k}$) and sends the trap door set $\{T_{w'}\}_{w'\in S_{w,d}}$ to the cloud server as the search request; At the same time, the image owner needs to temporarily save the trap door set $\{T_{w'}\}_{w'\in S_{w,d}}$, which will be used to verify the search results.

(2) Search ($G_W, \{T_{w'}\}$): this algorithm is executed by the personal cloud server in two cases: if the search is successful, the personal cloud server returns the corresponding document address and search evidence to the user. If the search fails, the personal cloud server returns the corresponding search evidence to the user.

1)After receiving the search request from the image owner, the personal cloud server first divides each trap door into l/n parts, and each part is represented by the corresponding letter in a Δ , so that each trap door is converted into a corresponding symbol sequence.

2) The depth-first search is performed for the cloud service on the index tree by using the corresponding algorithm to return the document address and search evidence to the image owner.

(3) Verify ($k, proof$): the core idea of validation is to search for a globally unique value for each nodal r_1 in the tree, called $proof$ .The specific algorithm is as follows:

1) Case of successful search:

A. The data owner verifies that the data of $gk(I\hat{D}_w)gk(ID_w)$is valid. If so, proceeding to the next step; Otherwise, the output false indicates that the result returned by the server is incorrect;

B. the data owner verifies that the data of $F_k(path\,||\,\hat{}\ \mathrm{ID}_{w_i}) = F_k(path\|ID_{w_i})$ is valid. If so, the output is true; otherwise, the output is false.

2)Case of unsuccessful search:

A. The data owner verifies that the data of $F_k(path\,||\,\hat{}\ \mathrm{ID}_{w_i}) = F_k(path\|mem)$ is valid. If so, proceeding to the next step; Otherwise, the output false indicates that the result returned by the server is incorrect;

B. the data owner verifies that the data of $mem[ord(T_w[i+1])] = 1$is valid. If so, the output is true; otherwise, the output is false.

3.3. The Hash function

The Hash function through the Hash algorithm, can transform input of any length into output of a fixed length. The value of the output is called Hash value or message digests. The Hash value can be expressed in function H:

$$h = H(M)$$

Where: M is a message of variable length, h is a message digest of fixed length.

SHA-3 is a Hash function family based on sponge function, which is obtained after adding special filling algorithm; defined as $Keccak[r, c]$.The sponge function can receive input of any length and satisfy output of any length. It is composed of three parts and has the following structural features: 1) internal state $S = R \| C$ and length is $b = r + c$;The parameter r is called the conversion rate and c is called capacity.2) permutation function f ;3) message padding, making the input length is multiple of the integer about r .The operating mode of sponge function can be divided into "inhalation" and "exhalation" operations, as shown below:

(1) Inhalation of sponge function

1). Initialization $S = 0$.

2). Fill the message and divide the message into data blocks for length of r ;

3). Message block and R go to XOR ;

4). S convert to $f(S)$ through the displacement function f ;

5). Let $S = f(S)$, iterating until all message blocks are processed.

(2) Suction of sponge function

1). The first bit about r of internal state S after the output iteration is completed;

2). If more output is needed, replacing S with $f(S)$ and putting out first bit of r about $f(S)$;

3). Let $S = f(S)$, repeating the above steps until the output length meets the requirements. If the length of the message digest is not an integer multiple, truncating it.

The width of the permutation function $Keccak - f(b)$ used in Keccak is determined by $b = r + c$ .Where: $b \in \{25,50,100,200,400,800,1600\}$(Fixed width is 1600 in the SHA-3 standard) .The algorithm details are as follows:

(1) Formal bit - Reordering rule

Within the Keccak, inputting information M follows a low to high order of significant bits.That is: the rules of information processing before entering the algorithm.

(2) Multi-ratep-pad01 *1 fill rule

At the end of the message M , filling a 1 first, then a number of 0, and finally a 1;Where: the number of 0 is an integer multiple of the length of r about the filled message.Let's call k , the fill length is at least 2 bits, at most r+1.The number of zeros filled in is $kr-|M|-2$.

(3) Permutation function $Keccak-f(b)$

$Keccak-f(b)$ is the iterative permutation function, is the basic core function of Keccak algorithm. According to different values of permutation width b, there are 7 different iterative permutations, number of iterations is $n_r=12+2l$, where: $2^l=b/25$

After the public key searchable encryption scheme, verifiable fuzzy keyword search scheme and SHA-3 standard Hash function are constructed, the ciphertext image remove duplicate according to the flow chart of ciphertext image. In order to improve the efficiency of the algorithm, image owners can classify the image based on its color feature, texture, shape and spatial relationship, and adopt different private key encryption according to different categories. For the message digest h_1 of the image x on the original random key k, when $h\neq h_1$,encrypting the newly generated random key k_1 and the random number k obtained in the past, together with the message digest h_1, h and ciphertext index b, are outsourced and stored on the personal cloud server, so that the ciphertext image can be deleted duplication when uploading next time.

4. Conclusion

In this paper, a cryptographic image deduplication method based on public key searchable cryptographic technology is implemented by using personal cloud server with good security and service quality. A verifiable fuzzy keyword search scheme with the advantages of reducing the computation and verification cost of the client is used to search the search data. Then The high efficiency of searching data is achieved. Due to the anti-collision of Hash function, by comparing the message digest of ciphertext image, the accuracy is improved. And the public key searchable encryption technology based on the EIGmamal public key algorithm can not only realize the security encryption of the image, but also meet the private conversations between people who do not know each other.

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).

References

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

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

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

iv. *Gai, K., Qiu, M., & Zhao, H. (2018). Energy-aware task assignment for mobile cyber-enabled applications in heterogeneous cloud computing. Journal of Parallel and Distributed Computing, 111, 126-135.*

v. *Yuan, Haoran, et al. "DedupDUM: Secure and scalable data deduplication with dynamic user management." Information Sciences, 456 (2018): 159-173.*

vi. *Jiang, T., Chen, X., Wu, Q., Ma, J., Susilo, W., & Lou, W. (2017). Secure and efficient cloud data deduplication with randomized tag. IEEE Transactions on Information Forensics and Security, 12(3), 532-543.*

vii. *Shin, Y., Koo, D., & Hur, J. (2017). A survey of secure data deduplication schemes for cloud storage systems. ACM Computing Surveys (CSUR), 49(4), 74.*

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

ix. *viiii. Zhang, Y., Feng, D., Jiang, H., Xia, W., Fu, M., Huang, F., & Zhou, Y. (2017). A fast asymmetric extremum content defined chunking algorithm for data deduplication in backup storage systems. IEEE Transactions on Computers, 66(2), 199-211.*

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

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

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