Available online www.jsaer.com

Journal of Scientific and Engineering Research, 2018, 5(9):215-220

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Secure Deduplication Scheme for Cipher text Images

Jun Ye, Jianqiang Xu, Jike Wan, Yu Cao, Zouyu Xie, Liufen Li*

School of mathematics and statistics, Sichuan University of Science and Engineering, Zigong, Sichuan, China

Abstract With the rapid development of cloud computing and cloud services, more and more enterprises and individual users upload their local images to the cloud to reduce local storage space. And the cloud server is not completely trusted, it may steal users' information. Therefore, users should encrypt the image before uploading it to protect the privacy of user information. Aiming at this problem, this paper adopts the method of ecntypted image deduplication. The key is obtained by calculating the original image using MD5 hash encryption algorithm. The CR5 encryption algorithm is used to encrypt the original image to obtain an encrypted image, that is, the image that the user will upload. The hash algorithm is used again to obtain the hash value of the ciphertext. In this way, the accurate elimination of duplication can be achieved, and the information of users can keep secret.

Keywords Image deduplication, Security, Hash function

1. Introduction

Cloud technology [1] is a huge boon to consumers: it allows them to store vast amounts of information at low or no cost such as music, information, photos, and so on. With a variety of services, people can store more data at will without buying additional devices, such as hard drives or memory sticks.

But with the rapid development of computer industry and microelectronics industry in recent years, popular mobile phone cameras market now is generally 12 megapixels, and the size of the 12 megapixels photographs about 4M, as the picture size is generally promoted, the problem of insufficient storage space in the cloud become a headache for different cloud server companies. How do you store the most users' data in the smallest space? First, the storage space of the company's cloud server is improved, but this is of little significance. If the company only considers using this point, when the user stores a little, the company will install a server, which will not only increase the maintenance cost, but also be inconvenient to manage, and the user is not limited to upload amount of data, which will form a serious vicious circle. Second, the company limits the upload data amount of users, such as setting the level. The higher the level, the more the upload data amount, which can control the problem of insufficient storage space in cloud to some extent. Third, image deduplication is to ensure that the same image does not exist in the server. By this point, the cloud can maximally save the storage space occupied by the image.

Although the cloud server is like a huge free storage cabinet, the storage cabinet is only a storage cabinet, it is not secure. It also has special security hidden trouble. Unlike physical filing cabinets, data stored in the cloud can be at risk of being stolen by global cybercriminals. If these criminals attack, they are likely to get a huge amount of information. Now data breaches have become one of the most common cyber security incidents in the world, and there is a growing trend. In the first half of 2017, 1.9 billion records were leaked or stolen globally, more than the total for all of 2016 (1.4 billion).

A prominent feature of the data breach in 2017 has been the expansion of the cloud data threat. With the maturity of cloud computing business model, more and more enterprises migrate their business to the cloud environment, and cloud database becomes more and more popular, which has a significant impact on the database security environment. Once the security vulnerability of cloud platform occurs or the security factors are not fully considered in the migration process, enterprise data may become the target of hackers.

On September 5, 2018, PICC Property and Casualty Company launched the "network information security insurance" product together with 360 enterprise security technology (Beijing) Company Limited and held a press conference. According to the blue book of cyber space security: Chinese cyber space security development report (2016) released by the information institute of Shanghai academy of social sciences and the security research institute of China academy of information and communication, the overall size of Chinese information security market is expected to reach 0.4822 billion dollars in 2019, which is about 3.3189 billion yuan. In May 2017, President trump signed an executive entitled "enhancing the cyber and critical infrastructure cybersecurity of the federal government", which optimizes and upgrades the top-level design of American cyber space security through adjustments to existing strategic policies. In July of the same year, Singapore published a draft of new cybersecurity regulations aimed at safeguarding national cybersecurity, maintaining critical infrastructure and empowering authorities to perform necessary regulatory duties. Through the enterprise cooperation, the implementation of the national legislative measures can be seen that both the enterprise and national or the Third international. Top-level network security design has been from the slow lane in the past into the fast lane, article from coarse to detailed, the future will promote the modern design of network security management system is all over the world, all the mainstream trend and development of the enterprise needs.

Therefore, the problem of information security is extremely urgent. When we upload image to the cloud, we should consciously pay attention to the security of image information we involve in privacy. The March 2017 Hollywood star nude photo scandal. It is because the victims do not pay attention to the security of their private images, so that criminals can take advantage of it, resulting in privacy leakage. Therefore, information security should be considered more in the process of image deduplication.

1.1 Related work

With the rapid development of the network and the rapid popularization of cloud storage service, image, as a common material in People's Daily life, has become the focus of people's attention on how to save the local space occupied by these trivial files, which has triggered people's attention to the repeated data and information security [2] in the cloud server. In recent years, many effective solutions have been proposed to improve the utilization rate of cloud server space, save network bandwidth and secure users' information. Started research direction is to delete existing data in the server side [3], but not flexibly access data and control users' rights and exist the condition of the mistaken delete [4], for example, the current image is doing one of the common type of shared data in cloud storage, image data in cloud is easy to reach the level of a server in a short period of time is difficult to deal with, it not only greatly wastes of the storage resources, but also affects users, because the image similarity, the probability of image deletion error is very high, so people introduce the concept of ownership [5], and adding in the access library into the server for users' data, therefore, users need to submit their own permission certificates to the server when extracting data. In this way, it not only reduces the probability of false deletion of users' data, but also guarantees the users' data security to a certain extent. Moreover, it explores various retrieval technologies, classification technologies and compression technologies reduce the time required for the retrieval of massive data in the server and optimize the users' experience. And as time goes on, the number of cloud server users grows, and the amount of data that needs to be uploaded grows, requiring a method to save bandwidth better. People explores a more effective way - the client deduplication, using the file tag to validation whether file is repetitive, reducing the communication bandwidth of uploading the duplicate files, but this way of checking repetitive image during the course of verification was found security is low, the hackers need only capture a small portion of the users' information, users' file labels or other data can access both users' uploaded data, in order to ensure the safety of the users' information, multiple encryption [6] methods are put forward to ensure client is checking repetitive image at the same time to ensure the safety of users' information of the client, but the effect is unsatisfactory. Recently developed a kind of fuzzily remove duplication [7] way, using some algorithms [8] to image by fuzzily repeatedly checking, then evaluate the image quality and choose the best quality image [9] saved in the server, for this way can improve greatly the information redundancy eliminating rate, but generally has some problems, such as a picture of the similar to keep those found, whether consistence’s with the users' goals, etc. Image of fuzzily remove duplication is the future development direction, but the technology is not very mature at present, we should further explore. In this paper, the image is deduplicated and some encryption methods are improved to make it more difficult for users' information to be leaked. At the same time, the bandwidth of image upload is further reduced and users' experience is improved.

2. Preliminaries

2.1 Symmetric key encryption

Symmetric key encryption is also known as private key encryption, that is, for plaintext encryption and decryption operations all use the same key. Symmetric encryption needs to be satisfied:

1) the analyst knows the algorithm and accesses some or more ciphertext, and cannot translate the ciphertext or get the key;

2) the security of the key. So as to ensure the confidentiality of plaintext.

2.2 Hash algorithm

Hash algorithm, also known as hash function, maps any length of binary to a shorter fixed length of binary value, which is called hash value. Hash value is a unique and extremely compact numerical representation of a piece of data, different sizes of files, under the same hash algorithm get the same hash value length. If a plaintext is a little different, it will produce different values after hashing. Hash algorithm has collision resistance, that is, it is very difficult to find different plaintext with the same hash value for a given plaintext; and it is very difficult to find different plaintext with the same hash value. Hash algorithm is commonly used for fast lookup and encryption algorithms.

2.3 RC5 algorithm

RC5 block cipher algorithm was invented by Professor Ronald L. Rivest of Massachusetts Institute of Technology in 1994 and analyzed by RSA laboratory. It is a block cipher algorithm with variable parameters. Three variable parameters are: block size, key size and encryption rounds. In this algorithm, three operations are used: XOR, add and loop. RC5 algorithm is highly efficient, needs small storage space, pays attention to security, and is favored by most cryptographers.

2.4. MD5 algorithm

Designed by American cryptographer Ronald Linn Rivest, MD5 is a widely used cryptographic hash function that generates a 128-bit (16-byte) hash value to ensure complete and consistent transmission of information and is widely used in information security.

3. Our scheme

Method applied in this paper of image removal duplication is image exact deduplication, encrypting on the original image by the MD5 hash encryption algorithm [10] and CR5 encryption algorithm [11], and then employing the MD5 hash encryption algorithm to encrypted image to get a safe and irreversible hash value, by uploading the hash value, let the cloud to find whether have the same hash value existing. Treatment scheme is:

(1) If the same hash value exists, the client will be feedback that the image does not need to be uploaded.

(2) If the hash value does not exist, upload the encrypted image and the hash value.

So as to achieve the goal of de-duplication.

3.1 Symbols

I_1 represents the local image that the user wants to upload to the cloud.

I_2 represents the same image on the cloud as the image transmitted by the user.

$h(I)$ represents the hash value calculated by the MD5 code hash algorithm for image I .

$E_k(x)$ means to encrypt x by RC5 symmetric encryption algorithm through the key h .

3.2 Image de duplication algorithm

3.2.1 Image removal in plaintext database

Assuming that the cloud is trusted, it will not divulge and destroy users' information and files. Then Image removal in plaintext, which can be divided into two steps:

(1) the user uploads the local image I_1 to the cloud, and the cloud determines whether the same image exists in the database.

(2) If the cloud does not have the same image as the image, it receives the image and feeds back information to remind the user that the image was saved successfully; otherwise, the cloud deletes the uploaded image and feeds back information to remind the user that the image is already on the cloud.

When users want to get their own uploaded images, they can send messages through the client to let the cloud feedback the images. The system model of plaintext image deduplication is shown in Fig.1.

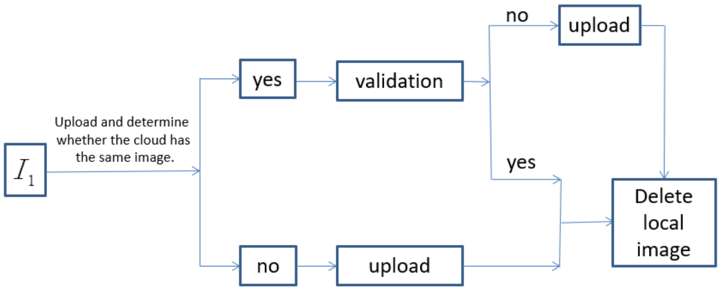

Figure 1: Deduplication model of plaintext database

3.2.2 Image deduplication in ciphertext database

In real life, the image de-duplication in plain text is not practical. There are two main factors: firstly, it is difficult for users to trust the third-party software, and users want to ensure their image privacy while also want to save it; secondly, the image is relatively large, uploading not only takes too much time but also consumes traffic, and users use and Inconvenient. To solve these problems, we will use the image de duplication algorithm in ciphertext.

(1) Plaintext is encrypted and transmitted by RC5, thus ensuring the security of user privacy.

(2) Using MD5 hash algorithm, the larger image is mapped to a very short binary value, which greatly reduces the transmission time and traffic of the client.

For image de-duplication under this ciphertext, the cloud will save the ciphertext corresponding to a large number of images and its hash value $\left(E_{h(I_i)}(I_i),\ h(E_{h(I_i)}(I_i))\right)$, when the user uploads the image to the cloud, the algorithm is shown in Fig. 2.

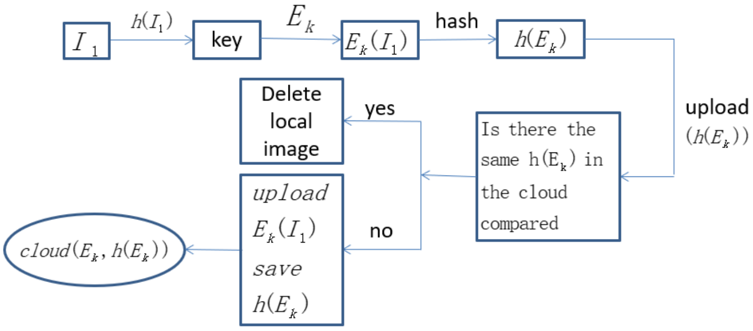

Figure 2: Deduplication algorithm in ciphertext database

(1) The client processes the image I_1 to be uploaded by MD5 hash algorithm, and obtains its hash value $h(I_1)$ and saves it locally.

(2) After encrypting the image I_1 by RC5 symmetric encryption algorithm, the ciphertext $E_{h(I_1)}(I_1)$ is obtained with the key $h(I_1)$.

(3) $E_{h(I_1)}(I_1)$ uses MD5 code hash algorithm to get its hash value $h(E_{h(I_1)}(I_1))$.

(4) Upload the resulting $h(E_{h(I_1)}(I_1))$ to the cloud and determine whether the hash value exists in the cloud database.。

(5) If the cloud does not have this hash value, it is necessary to continue uploading $E_{h(I_1)}(I_1)$ to the cloud as the image eigenvalue to save, and remind the customer to save successfully. Instead, the customer uploads information is deleted and the customer is reminded that the cloud already has the image.

Compared with plaintext image de-duplication, ciphertext image de-duplication increases the user's operation space and complexity to a certain extent. For users who do not know this knowledge and the algorithm process, it can cause problems when used. But on the other hand, image de-duplication under ciphertext pays more attention to the security of privacy, and becomes the gospel of most customers. In data transmission, the de-duplication of ciphertext greatly reduces the user's transmission time, the user gets the image shorter time, making the user more comfortable to use.

4. Conclusion

Image de-duplication is a method used by cloud service providers to reduce a large number of identical images. With the rapid development of cloud computing and cloud services, more and more enterprises and individual users upload their local images to the cloud to reduce the possession of local memory of images and facilitate them to download the cloud images on different ports, resulting in a large number of the same redundant images generated in the cloud.

In order to solve this problem, in the first time, people come up with that through the client to upload their own target image to the cloud to ask if there is the same image, the cloud compares target image with image in the cloud to determine whether the cloud has the same image as target image, if it has, then saving the file name in the cloud and making file link to the existing same image, if not, file name and the target image store in the cloud, and the file name links the image. Each time the client uploads an image, it goes through the process to remove duplication. However, this method can cause the users' privacy to be stolen, modified or lost, so the de-duplication in clear text is not suitable for the current environment and needs. The method of removing duplication under ciphertext also comes into being. According to the market demand, the fuzzily removing duplication image method appears. This method compares similar images and keeps the image with higher quality and better pixel. However, this method also creates a problem, because the image distortion is so serious that the user loses important information, and at the same time, some events that delete important images of the user by mistake will make the service provider get into trouble.

This paper adopts the method of accurate ciphertext image de-duplication to solve this problem. On the basis of encryption and de-duplication, the image integrity is also ensured. Several identical images are removed to ensure that the image is not distorted. In this paper, the key obtained from the calculation of the original image using MD5 hash encryption algorithm is known to be unbreakable, that is, the key is very secure and uncrackable [12]. Then through CR5 encryption algorithm to encrypt original image to get an encrypted image, which the user want to upload, using a hash algorithm to the encrypted image to get the hash value of the encrypted image again, using the hash value to judging whether the cloud has the same image with the encrypted image going to be uploaded, so as to achieve the aim of removing duplication precisely, and can protect the users' information from the cloud knows, also don't have to worry about stealing images is a third party.

However guaranteeing information security still needs better ways to support, the situation of the cloud is completely unreliable can't be solved, for example, when users upload the encrypted image but cannot know whether the cloud existing the same encryption image as uploaded image, therefore we need to solve the problem of the client verifies if the cloud has the same encryption image, that would rule out the possibility that the cloud of cheating.

Acknowledgements

This work was supported by 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); the key laboratory of higher education of Sichuan Province for enterprise informationalization and Internet of things (No.2017WYJ02).

References

[1]. Ye, Y., Hu, T., Zhang, C., & Luo, W. (2018). Design and development of a CNC machining process knowledge base using cloud technology. *The International Journal of Advanced Manufacturing Technology*, *94*(9-12), 3413-3425.

[2]. Fernandes, D. A., Soares, L. F., Gomes, J. V., et al. (2014). Security issues in cloud environments: a survey. *International Journal of Information Security*, *13*(2), 113-170.

[3]. Meyer, D. T., & Bolosky, W. J. (2012). A study of practical deduplication. *ACM Transactions on Storage (TOS)*, *7*(4), 14.

[4]. Ren, Z., Wang, L., Wang, Q., & Xu, M. (2018). Dynamic proofs of retrievability for coded cloud storage systems. *IEEE Transactions on Services Computing*, *11*(4), 685-698.

[5]. Yang, C., Ren, J., & Ma, J. (2015). Provable ownership of files in deduplication cloud storage. *Security and Communication Networks*, *8*(14), 2457-2468.

[6]. Zheng, Y., Yuan, X., Wang, X., Jiang, J., Wang, C., & Gui, X. (2017). Toward encrypted cloud media center with secure deduplication. *IEEE Transactions on Multimedia*, *19*(2), 251-265.

[7]. Ma, K., Duanmu, Z., Wu, Q., Wang, Z., Yong, H., Li, H., & Zhang, L. (2017). Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, *26*(2), 1004-1016.

[8]. Joly, A., Buisson, O., & Frélicot, C. (2007). Content-based copy retrieval using distortion-based probabilistic similarity search. *IEEE Transactions on Multimedia*, *9*(2), 293-306.

[9]. Sheikh, H. R., Sabir, M. F., & Bovik, A. C. (2006). A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, *15*(11), 3440-3451.

[10]. Binti Suhaili, S., & Watanabe, T. (2017). High-Throughput Message Digest (MD5) Design and Simulation-Based Power Estimation Using Unfolding Transformation. *Journal of Signal Processing*, *21*(6), 233-238.

[11]. Kuznetsov, A. A. (2015). On the cryptographic security of the “BotikKey” authentication protocol against attacks on MD5 hash function. *Program Systems: Theory and Applications*, *6*(1), 135-145.

[12]. Sharma, D., Sarao, P., & Dudi, S. (2015). Implementation of Md5-640 Bits Algorithm. *International Journal of Advance Research in Computer Science and Management Studies*, *3*(5), 286-293.