Research on a Cipher-Text Fuzzy Retrieval Algorithm Based on Spatial Model

Yuhua Wei

Guangzhou Huali Science and Technology Vocational College, Guangzhou 511325, China.

Abstract. with the development of the Internet, cloud computing and cloud storage are more and more widely used. Users are starting to store more and more large data files on cloud servers. In the cloud storage mode, some private data and sensitive data of users may be leaked. To protect data security, users need to encrypt their private data before storing it in the cloud server. However, when viewing the data, the user needs to restore the encrypted data, and then go back to the original search, which is a waste of time. Therefore, it is of great significance to study how to effectively and quickly realize cipher-text data retrieval [1]. When the data scale is large, the traditional retrieval structure is often inefficient and inaccurate. Therefore, how to ensure the security of cipher-text retrieval and how to improve the efficiency and accuracy of keyword fuzzy retrieval has become the focus of experts and scholars. Because of the low efficiency of cipher-text retrieval, combined with the current cipher-text retrieval scheme, a fuzzy retrieval method for cipher-text with multiple key words is proposed. This method is aimed at cipher-text retrieval, which realizes fuzzy retrieval of multiple keywords. In addition, it can also resist the trap door association attack, adaptive keyword selection attack and similarity attack.

Keywords: cloud storage cipher-text retrieval multiple keywords fuzzy retrieval.

1. Introduction

Due to the development of the Internet and cloud computing technology, more and more individuals and companies begin to store large amounts of data files in cloud servers. This saves storage space and provides quality service. As one of the core technologies of cloud computing, cloud storage technology has been paid much attention by researchers. However, as a third party server, the cloud server will perform corresponding operations according to the relevant protocol of the system, that is, it will actively detect the file content stored on it. When data is uploaded on the cloud server, the user cannot directly control these data, thereby increasing the risk of privacy, such as id card number, telephone number and personal E-mail information [2].

Therefore, effective data protection measures and privacy protection technologies are of vital importance, especially when it involves some private data of users or some important data information of government departments. In fact, there have been frequent data leaks in recent years, resulting in a large number of user privacy leaks. In order to make customer privacy data more secure, prevent was attacked by illegal user, the user will be after the individual privacy data encryption, and upload the encrypted data to the cloud server, can ensure the privacy of sensitive data in the transmission process. However, encrypted data has lost some features of plaintext data, which makes it difficult to retrieve these data efficiently [3].

The traditional information retrieval method is not suitable for cipher-text data retrieval in cloud environment. In order to solve the cipher-text retrieval problem in the cloud server, domestic and overseas researchers have conducted an in-depth study. The existing schemes are divided into two categories: precise keyword retrieval and fuzzy keyword retrieval. First generally is the process of the data owner to key words indexing, to store the index encrypted to cloud server, after the authorization by entering an encrypted data users query keyword search.

Although the existing cipher-text retrieval technology can ensure the efficiency, security and accuracy of the retrieval process. However, there are still some shortcomings in these plans. First, the existing cipher-text retrieval technology only supports single keyword retrieval. With the continuous expansion of data scale, limited keywords cannot accurately describe users' needs, so users need to conduct multiple keyword search for cipher-text. Secondly, most of the existing cipher-text retrieval techniques only allow precise keyword query. The user cannot perform fuzzy retrieval. In addition, most cipher-text retrieval technologies do not sort the retrieval results, causing the client to deal with a large amount of encrypted data. In conclusion, it is extremely necessary to study and improve the existing cipher-text retrieval technology.

Current typical cipher-text retrieval scheme includes the following: linear full-text search algorithm, the security index algorithm, introducing the sort of encryption algorithm, keyword search algorithm based on the male pin encryption, the retrieval algorithm based on the homomorphic, cipher text fuzzy retrieval algorithm. However, there are some problems in these retrieval schemes, such as low efficiency, large space storage, no support for multi-keyword fuzzy retrieval and no support for retrieval result ordering [4]. Therefore, in order to solve the above problems, on the basis of the above research, this paper proposes a multi-keyword cipher-text fuzzy retrieval.

2. Overall Framework of the System

2.1 System Model Analysis

In this paper, we study the multiple keywords fuzzy retrieval system for cipher model the overall structure is mainly composed of three parts: The Data Owner, authorized Data Users and Cloud Server. As shown in figure 1, data owners upload data files and index files to the cloud server after the client encrypts them. The authorized data user submits the key word Trapdoor to the cloud server; The server sorts the matching results and the similarity between the data files and returns the matching results to the authorized person.

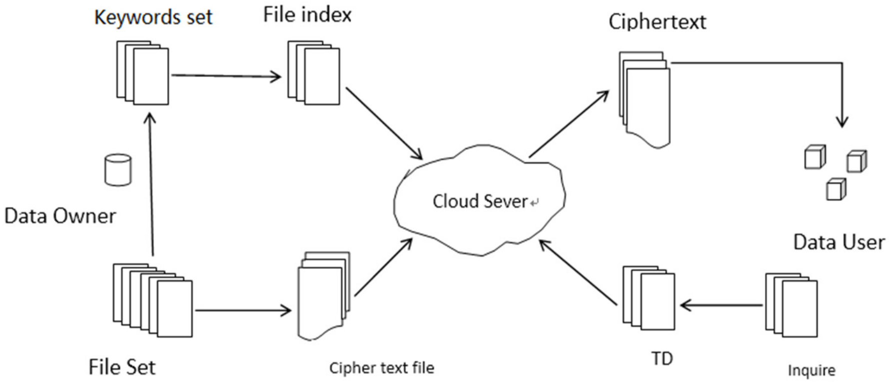

Figure 1. structure diagram of multi-keyword retrieval system model

2.2 Cipher-text Retrieval Process

In order to facilitate the study of cipher-text retrieval process in this paper, the system model assumes that the authorization between data owner and data user has been completed. The basic process is shown in figure 2 below.

Tell from the process, keyword oriented cipher-text retrieval method can be roughly divided into four stages, including the encryption Key generation, Build Index, Trapdoor and Match the keywords. The specific contents of these four stages include: vector Index building, vector Index building, Trapdoor searching and keyword matching [5].

Key Gen: first, the data owner enters a random number according to the selected data encryption scheme, thus generating the encryption Key. The data owner then encrypts the keyword index structure and data files before uploading. Usually the alternative encryption schemes are symmetric encryption schemes and asymmetric encryption schemes. This paper chooses symmetric encryption scheme according to its own characteristics and needs.

Vector Index: the data owner processes the data file C using a keyword extraction algorithm. According to the mapping relationship, the data owner can generate the plaintext keyword index structure I. Then, the index I is encrypted according to the encryption key in the previous step, and the keyword encryption index I 'is generated, and the index I' is submitted to the server together with the encrypted data file.

Search trap generation (Trapdoor): the search keyword submitted by the authorized user can be represented as Query = {(w'1, w'2, ..., w's) |w'i∈ W, 1 ≤ i ≤ s ≤ l}. Local structure of vector generated keywords in the same way, and USES the symmetric encryption algorithm to encrypt the keyword vector structure, according to the encryption strategy generated query keywords trapdoor T omega.

Keywords matching (Search): the data user submitted the query of tropdoor T axial to the cloud server. The cloud server matches the keyword index and trap structure, and calculates the similarity score. After sorting according to the score, the most relevant encryption file of top-k is given to the data user. Finally, the user downloads the files and decrypts them.

In the four stages above, the generation of encryption keys and the establishment of the data file keyword index structure usually require only one run by the data owner to upload to the server. And authorized users not only need to submit their data retrieval requests have cloud service to cooperate with the keyword matching, namely the need to complete the keywords security trapdoor generated and query two steps. In other words, in a cloud environment, users are typically "thin client" type, which is to minimize the workload on the client side. In conclusion, in order to reduce the working intensity of the client, in the framework of cipher-text retrieval, it is necessary to simplify the calculation of trap door generation and keyword retrieval in the client as much as possible. The cloud server should handle as much computing as possible while ensuring security and efficiency.

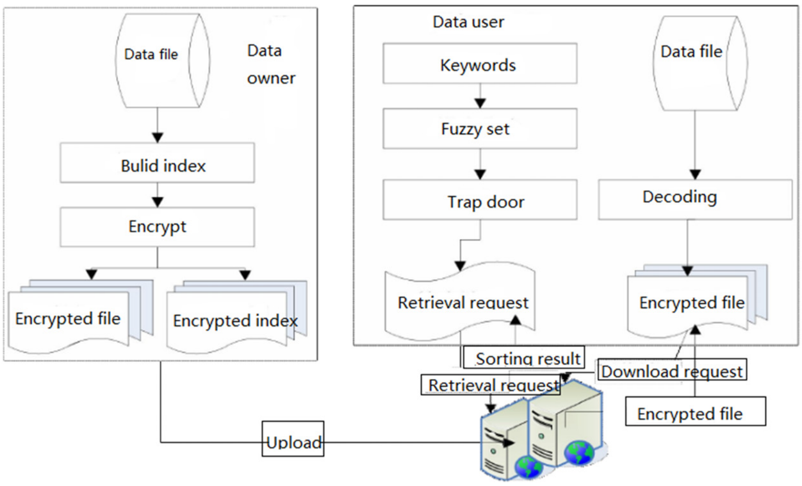

Figure 2. Basic Process diagram

3. Scheme Design

3.1 Program Flow

When it comes to multi-keyword fuzzy retrieval in encrypted files, we should improve the original bit-array structure of Bloom-Filters. We can combine the Locality - Sensitive Hashing (LSH) functions of keyword vector structure mapping, a new data structure is constructed to represent the more keyword vector structure, finally USES the invertible matrix encryption algorithm to encrypt the keyword vector structure. The scheme on the basis of security KNN retrieval technology, not only can realize multiple keywords fuzzy retrieval, but also can effectively resist the trapdoor correlation attacks, such as simplifying the keyword index of encryption and decryption process, optimize the keywords of cryptograph query efficiency. The main implementation process of this program is shown in figure 3 below.

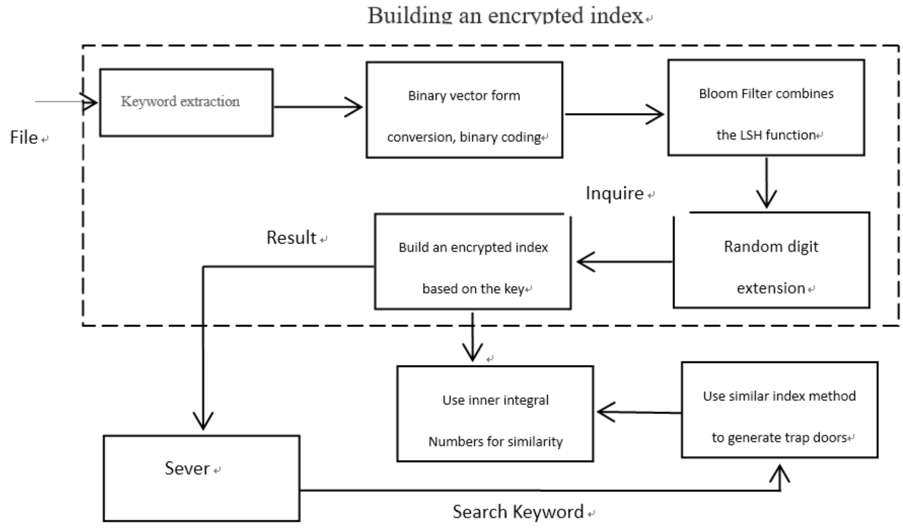
Building an encrypted index
File
Keyword extraction
Binary vector form conversion, binary coding
Bloom Filter combines the LSH function
Inquire
Result
Build an encrypted index based on the key
Random digit extension
Use inner integral Numbers for similarity
Use similar index method to generate trap doors
Sever
Search Keyword

Figure 3. The program execution process

3.2 Locally Sensitive Bloom Filter(LSBF)

The traditional Bloom Filter USES the standard hash function to perform hash mapping on all elements in the collection, generating bit array structure. But because of the limitations of the standard hash function, we cannot hash elements with relatively small edit distances to the same place. That is, fuzzy retrieval of multiple keywords cannot be supported under cipher-text conditions [6]. Therefore, the locally sensitive hash function is used to replace the standard hash function. In addition, it can be concluded from the analysis in the previous section that the standard Bloom filter will generate false positive cases and the locally sensitive hash function will generate additional false negative cases. In order to ensure the accuracy of fuzzy retrieval of multiple keywords, this paper improves the vector structure of Bloom Filter and keywords on the basis of ensuring the accuracy of keyword cipher-text retrieval.

Locally sensitive Bloom Filter refers to selecting L locally sensitive hash functions independently of each other, gi(1≤i≤L), and then mapping elements to m-bit bitwise arrays. It is used to replace the locally sensitive hash bucket structure, thus reducing space coverage. When each element in the collection is mapped to a bit array by the hash function gi(1≤i≤L), the number of digits generated during the mapping can be considered as the vector structure of the collection space. This spatial vector structure supports fuzzy retrieval. Specifically, when a user submits a retrieval request q, the same l locally sensitive hash function gi(1≤i≤L) is used in this article. It then requests the set to be processed and determines the similarity between the request set and the original set by calculating the number of internal integrals of the two vector structures. The design of locally sensitive Bloom Filter is shown in figure 4.

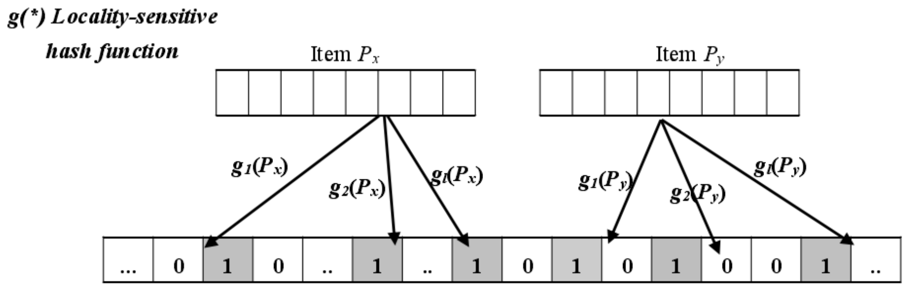

Figure 4. design of locally sensitive Bloom Filter

3.3 Keyword Vector Structure Security Bit Expansion

In this design, LSH function is used to replace the conventional hash function in Bloom Filter. According to the hash process of LSH function, the input variable of hash is vector form. So when we map using the binary encoding of the keywords, the index construct is shown in figure 5.

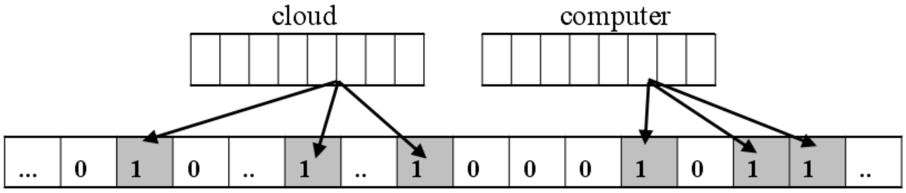

Figure 5. Bloom Filter index construction

4. Result Analysis

This paper defines the similarity correlation between keywords and data files, and analyzes the security of cipher-text retrieval. The security of cipher-text retrieval scheme is normalized for the first time. This paper proves that the traditional server side hierarchical cipher-text retrieval scheme based on sequential encryption will inevitably leak the characteristic values between private data and data. So we need the original cipher-text retrieval algorithm was improved, the guarantee of keywords retrieval efficiency at the same time, to ensure the privacy of data files, avoid credibility to the cloud server to leak privacy.

References

[1]. Wang Linsong, Liu Deshan, Guo Jin, et al. Public cloud security architecture design [J]. Journal of Jilin university: information science edition, 2013(2):165-169.

[2]. Xie Xianming. Research on ciphertext retrieval technology based on search history [D]. National university of defense science and technology, 2011.

[3]. Wang C, Chow S M, Wang Q, et al. Privacy-Preserving Public Auditing for Secure Cloud Storage[J]. IEEE Transactions on Computers, 2013, 62(2):362-375.

[4]. Huang Yongfeng, Zhang Jiuling, Li Xing. Encryption storage and retrieval technology in cloud storage applications [J]. Zhongxing communications technology, 2010, 16(4):33-35.

[5]. Guo Lulu, Xu Chungen. Research on cloud storage ciphertext retrieval methods [J]. Information network security, 2013(9):6-9.

[6]. Song Yan, Han Zhen, Chen Dong, etc. Attribute encryption scheme supporting keyword random connection search [J]. Journal of communication, 2016,37(8):77-85.

[7]. Collaboration a. Search for the b(b)over-bar decay of the Standard Model Higgs boson in associated(W/Z)H production with the ATLAS detector[J]. Journal of High Energy Physics,2015 (1):1-89.