Encryption Cipher Text Retrieval Scheme Based on Fully Homomorphic Encryption Enterprise Cloud Storage

Lijuan Wang¹, Lina Ge², Bo Geng³, Qiuyue Wang⁴

Key Laboratory of Network Communication Engineering, School of information science and engineering, Guangxi University for Nationalities, Nanning 530006, China

18377149170@163.com

Abstract: With the development of computer networks, cloud storage has become the mainstream way for people to store information. However, some essential information is disclosed during the storage process and the security of information has become the most concerned issue at present. In order to use the powerful computing of cloud storage to realize resource sharing between relevant departments without revealing information of other unrelated departments, this paper proposes an encryption cipher text retrieval scheme based on fully homomorphic encryption enterprise cloud storage (ECRS), and designs an Enterprise-side security model. The model uses full homomorphic encryption, decryption and spatial vector cipher text retrieval to achieve the security of information and resource sharing among enterprise departments. ECRS determines the employee who can encrypt or decrypt the files, and who has access to the files by modifying the number of attributes intersected between attribute sets. The solution can provide the security for cloud storage, resources for different departments, and improve the accuracy for cipher text retrieval.

1. Introduction

With the development of computer network, more and more individuals and enterprises rely on cloud storage due to the powerful storage capacity and computing capacity of cloud storage. Single data integration is stored in an open environment, where important information is easy to be stolen by illegal authorities without the protection of firewalls and other security measures, so people have higher and higher requirements on the security of information on the basis of relying on cloud storage. Enterprises can pay a certain fee to store a large amount of data in the cloud, which can improve their data storage and save some space. However, the security of data and its integrity still make users have doubts, and they will consider whether the data stored in the cloud by enterprises will be lost, which will lead to the loss of competition between enterprises. If you can’t meet the requirements of the above basic, even if the cloud storage has a powerful storage capacity and computing power, it is also useless for businesses and other users.

In this paper, cipher text policy attribute based encryption (CP-ABE) [1], which is based on cipher text rules, is combined with full-homomorphic encryption technology to encrypt and store enterprise secret data in the cloud, and relies on attributes for data access; in terms of data sharing and data confidentiality, multi-keyword retrieval technology is proposed to set up the basic security architecture of enterprise end. It solves the problems of confidentiality, integrity and sharing of stored data, and provides a retrievable scheme for data under secure transmission and access control.
2. Material and methods

2.1. Data encryption, access control and cipher text retrieval

With the development of computer, no matter storing or transferring data in the network, the data should be encrypted. The encryption of data can ensure its security, confidentiality and data integrity, and avoid the theft and modification of data by illegal users in the process of storage and transmission. The holomorphic encryption algorithm was proposed by Rivest et al. [2] in 1978, and later was called by researchers as the "holy grail"[3] in the field of cryptography. The characteristic of the algorithm is to calculate the cipher text without knowing the key. After decryption, it is equivalent to the corresponding operation on the plaintext, which satisfies the requirement that f(Enc(m))=Enc(f(m)). It was not until 2009 that Gentry et al. [4] put forward the first all-homomorphic encryption scheme (all-homomorphic encryption scheme based on ideal lattice over polynomial ring). In 2011, Coron et al. [5] improved Gentry's scheme and put forward the scheme that adding public keys to cipher text in the form of product could reduce the size of public keys, which the size of public key is $o(\lambda^7)$.

Access control is to control the access of legitimate users to protected network resources, prevent illegal users from illegally modifying protected data resources, prevent illegal users from illegally accessing data through certain permissions, so as to achieve the security and integrity of network resources. Role-based access control model (RBAC) [6] is a widely used access control model. The RBAC96 model is to directly cancel the connection between users and permissions, introduce the concept of roles, bind roles with permissions, and then assign them to users, so as to achieve the purpose of assigning corresponding permissions to users. It makes the permission management more convenient and thus reduces the complexity of the system. The server encrypts the data preliminarily, and then uses homomorphic encryption to process the data preliminarily and encrypt the vector.

Cipher text retrieval mainly includes two types. One is the retrieval based on security index, which is to index the key words in cipher text and find whether the key words exist in the index. Let one is based on cipher text scanning retrieval, the method is to find the keyword and each word in the cipher text match and confirm whether the keyword exists. Cipher text retrieval is a continuous improvement from simple single-keyword retrieval to search about multi-keyword, and from low precision to high precision. Therefore, with the development of simple cipher text retrieval, the research on multi-keyword retrieval should be carried out. Cao[7] proposed a retrieval scheme of multi-keyword sorting, which uses coordinate matching to calculate the correlation between keywords and documents. It uses the similarity of inner product between two keywords to judge the relevance between words. However, it uses Boolean type to return data. It cannot directly judge which document uses the same keyword more frequently and which document has a higher correlation, nor can it give more accurate data of the document.

Not only can ECRS multi-keyword retrieval reduce the time complexity during operation, but also it can return more accurate document number by calculating the degree of correlation between word vectors.

2.2. Vector Space Model

Vector Space Model (VSM) was proposed by Gerard Salton of Harvard University. Document $D$ can be regarded as general text, where $t$ is the feature item, which is the language unit and appears in the document and can basically represent the basic content of the document. Therefore, the text can be represented as $D=(t_1,t_2,...,t_n)$ and $n$ are the number of characteristic terms. The text can be represented not only by the eigenvector, but also by the weight of the eigenvector, such as the weight given to the feature term $w$, and the text can be expressed as $D=(w_1,w_2,...,w_n)$. The basic idea of VSM is to express document $D$ and the keyword $Q$ to be queried in the form of vectors. During the retrieval and matching process, we can directly calculate the similarity between the vectors in the document and the vectors to be queried. The higher the similarity is, the greater the connection between the two vectors will be, so that the accurate correlation between keywords and documents can be calculated, thus ensuring the accuracy of retrieval.
3. Storage framework design

3.1. The composition of the storage framework

In order to solve the problem that enterprises can safely store data in cloud storage, avoid illegal operations such as data theft and modification by illegal users, ensure the security and integrity of data, and propose a security model suitable for enterprise storage, see figure 1.

1) The first part is the company, which stores data in the cloud. There are several departments in the company, and each department employee gets his/her own property certificate through the property assignment server in the internal total server of the company.

2) The second part is cloud storage, which is responsible for storing homomorphic encrypted vector files. Calculate the correlation score between multiple retrieval vectors and document vectors to get the document number with high correlation score and return to the user.

3) The main operations of the total control server in the company are as follows:
Issue the employee's property certificate. It expands and encrypts plaintext vectors. Receives retrieved data and decrypts it, returning it to a specific employee.

![Figure 1. Storage model for data encryption, decryption and retrieval](image)

The security model is solved:

1) Data security

Only users of the enterprise itself and those who have been granted access rights can view and modify the data information, and no other users can access and modify the data.

2) Data integrity

In order to prevent data from being changed by unauthorized users or unauthorized data from being changed, this paper uses homomorphic encryption and decryption to process the data. The whole process of transmission and storage is in the form of cipher text, thus ensuring the integrity of data.

3) Data sharing

Relevant departments in the enterprise can consult relevant data information to achieve resource sharing between departments.

4) Effectively encrypted retrieval of data

The enterprise data is so huge that the word vector multi-keyword retrieval algorithm can be used to quickly obtain the data information with high accuracy under the condition that the effective encryption does not touch the plaintext information.

3.2. The process of data encryption and cipher text access in the storage framework

Employees in each department of the company have property assignment certificates issued by the server, which reflect the basic property information of employees. When an employee stores data in the cloud, he needs to submit the data together with the certificates he owns to the general server. After receiving the certificate, the server accesses the certificate and values the parameters used by the encryption algorithm according to the certificate content. Then, according to the initially encrypted document set $D_s$, document vector $D$ in the vector space model is generated. The public key $P_k$ given by the key generation algorithm of homomorphic encryption is used to encrypt $D$ by the full homomorphic encryption algorithm. Then it is uploaded to the cloud storage, so that the cloud storage service provider cannot know the specific content of the data, ensuring the confidentiality and integrity
In the traditional asymmetric encryption, we can only crack the encrypted data with a private key, so that the cipher text data cannot be shared. In this paper, Bethencourt et al. proposed the CP-ABE scheme [1]. The basic idea of this scheme is that the property sets of cipher text associations are different. When there are intersecting attributes between two attribute sets, and the number of attributes exceeds a certain value, the key associated with one attribute set can be used to decrypt the cipher text associated with the other attribute set. Just to meet the relevant departments can view each other’s relatively high correlation between the files, you can use your own key encryption departmental encryption files.

3.3. Weight of keywords and correlation score
In the field of information retrieval, multiple weight models are used to analyze documents. The basic rule: The more times a keyword appears in a document, the more important it becomes. If a keyword appears in many documents, it is not useful to distinguish between documents. If a document is particularly long and a keyword often appears, the keyword is not useful to use that keyword to distinguish documents. So how do you change that? In this paper, weight formula (1) is adopted as follows.

\[ W_{ij} = (f_{ij} \times \ln n^a) \times 10^a \]  

(a=lg(N/(ni)), \( w_j \) represents the weight of keyword \( I \) in document \( j \), \( f_{ij} \) represents the number of times the keyword \( i \) appears in \( j \). \( N \) represents the total number of documents in the document collection. \( N_i \) represents the number of documents that contain \( I \). \( e \) depends on the requirement of queue accuracy during retrieval).

Document vectors and query vectors have the same dimension \( n \). \( n \) corresponds to the total number of different eigenvectors in all documents. The correlation score between two vectors is proportional to the degree of matching between them. The correlation score is expressed by the inner product between vectors:

\[ D_i = (w_{1i}, w_{2i}, ..., w_{ni}) \]  

\[ Q_q = (w_{1q}, w_{2q}, ..., w_{nq}) \]  

\[ \text{Score} = \text{sim}(Q_q, D_i) = \sum_{k} w_{kq} \times w_{ki} \]  

In vector space, the similarity between two vectors is determined by the correlation score of two vectors \((D_i, Q_q)\). The correlation score is proportional to the similarity. The higher the correlation score, the two vectors are more similar. In the retrieval process, the correlation scores between two word vectors are used to retrieve similar files to achieve the optimal query results.

4. Full homomorphic encryption scheme steps

4.1. Value of parameters and Algorithm process of key generation:
Setup (\( \lambda, L \), \( \lambda \) is the set safety parameter, \( L \) refers to the number of circuit layers that are computed in homomorphic encryption, \( P_i \) represents the random number generated in each circuit layer. \( P_i \) also includes a decreasing modulus sequence \( Q_l = (l=L, L-1, ..., 0) \). The Gaussian Noise on an \( R \) satisfies the distribution of \( \xi \).

\[ \text{KeyG}(\{R\}): i = (L, L-1, ..., 1, 0) \]

Perform the following operations:
Setp1: get the public key \( Q_i \) of layer \( I \), the private key \( S_i \in R_q^2 \)
Setp2: calculate \( S_i \otimes S_i \rightarrow S_i' \)
Setp3: perform SwitchKey \( (S_i', S_i) \rightarrow \tau_{S_i} \rightarrow S_i' \)
Setp4: output the private key \( S_i = M + Q_i \rightarrow C \), the public key \( P_i = (Q_0, Q_1, \tau_{S_i} \rightarrow S^i) \)

4.2. Encryption algorithm procedure and Decryption algorithm procedure
Enc(P, P_i, M): \( M \in R^2 \).
The plaintext space is a ring \( R=Z[X]/(X^4+1) \), \( d=2' (t \in Z) \), \( R_q=R/qR \), clear
$M=(M,0)$, I’m going to randomly pick the vector $r$ in $R^n$. Cipher computing $M+QLr \rightarrow C$.

Dec($P, S, C$): If the cipher text is at level $i$, the private key is $S_i$. Plaintext output is $M(<C, S_i> mod Q)mod 2$.

### 4.3. Homomorphic encryption algorithm procedure

**Add($p_k, C_1, C_2$):** If $C_1$ and $C_2$ have the same decryption key, then $C_1+C_2 mod q_i \rightarrow C_3$. Then calculate the final cipher text $C_4$ = $Refresh(C_3, S_i' \rightarrow S_i, Q_i, Q_i)$

**Mult($p_k, C_1, C_2$):** If $C_1$ and $C_2$ have the same decryption key, then $C_1 \otimes C_2 mod Q_i \rightarrow C_3$. Then calculate the final cipher text $C_4$ = $Refresh(C_3, S_i' \rightarrow S_i, Q_i, Q_i)$.

### 4.4. Cipher text update operation

$Refresh(C_3, S_i' \rightarrow S_{i-1}, Q_i, Q_{i-1})$:

- Key exchange: $SwitchKey(s_i' \rightarrow S_{i-1}, Q_i, Q_{i-1}) \rightarrow C_1$
- D/a exchange: $Scale(C_1, Q_{i-1}) \rightarrow C_2$

Where the key exchange prerequisite is the same module ($Q_i$). Convert cipher text $C$ corresponding to key $S_i'$ into cipher text $C_1$ corresponding to key $S_{i-1}$. The precondition of Scale is the same key, which converts cipher text $C_1$ under module $Q_i$ into cipher text $C_2$ under module $Q_{i-1}$.

### 4.5. Correctness of algorithm

**Theorem 1 (Search R-LWE)** [8]: $R=Z_q[x]/(x^n+1)$, $n=2^k$, $k \equiv l$, $q=1 mod 2^a$, $a=R^n$ (where $m$ is an integer, representing the dimension of $a$), $a$ is chosen randomly and uniformly, $e \in R$ to obey the normal distribution of $\Psi_a$ mistake, If we know $b \in R$, and $b=a* e + e$, we can solve the problem of $s$ from $(a, b)$.

**Theorem 2 (Decision R-LWE)** [9]: Let’s say that $R=Z_q[x]/(x^n+1)$, $n=2^k$, $k \equiv l$, $q=1 mod 2^a$, $s=R^n$, $a \in R^n$, $e \in R$ for a normal distribution of $\Psi_a$ mistake, $b=a* e + e$, and $b \in R$, remember $\Lambda_s$, $\Psi$ for the distribution of $(a, b)$.

### 5. Safety performance analysis

By analyzing the feasibility of this scheme, a specific application experiment was simulated to illustrate: Different users send three files to the general server, and the server analyzes the data and gets to know the data table 1:

| The file number | User roles             | Text attributes contained in the file (e.g., sales management) |
|-----------------|------------------------|---------------------------------------------------------------|
| 1               | The sales department   | Y                                                             |
| 2               | The Marketing Department | Y                                                          |
| 3               | Development department | N                                                             |
| 4               | The general manager    | Y                                                             |

Encrypting the data for the file data in the table and then retrieving the keywords. The retrieved content is shown in table 2.

| Query keyword | The file number |
|---------------|-----------------|
| quality       |                 |
| management    | 1, 2, 4         |
| design        |                 |
| market        | 1, 2, 4         |
| Wholesale management | 1, 2, 4     |

When accessing the data, the Down algorithm needs to be called according to the user’s role after
the query, and the results are shown in table 3:

| User roles            | The file number can be decrypted |
|-----------------------|---------------------------------|
| Development department| -                               |
| The Marketing Department| 2                              |
| Marketing manager     | 2、4                            |
| The sales department  | 1                               |
| Development department| 3                               |

**5.1. Safety analysis and Performance analysis**

This paper adopts attribute-based access control and a scheme combining holomorphic encryption, decryption and retrieval. In this scheme, employees need to have the same properties as those contained in the cipher text file when viewing the cipher text. This prevents illegal employees from stealing and modifying the data. Not only does it ensure the security of data resources, but also ensures the integrity of data.

There are four typical schemes that can be implemented in keyword retrieval. Their performance with ECRS is shown in table 4:

| plan       | Start       | D/a exchange | Approximate eigenvector |
|------------|-------------|--------------|-------------------------|
| DGVB       | $O(\lambda^{14})$ |              |                         |
| GSW13      | $\tilde{O}(n \cdot (nL)^{\gamma})$ | $\tilde{O}(nL)^{\gamma}$ |             |
| Bra12      | $\tilde{O}(\lambda^{2})$ | $(\lambda^{3} \cdot L^{3})$ |             |
| BGV        | $\tilde{O}(\lambda^{2})$ |              |                         |
| ECRS       | $\tilde{O}(n)$ |              | $\tilde{O}(Nn)$         |

In the process of multi-keyword vector retrieval in ECRS, the inner product and similarity between vectors need to be calculated, that is, the inner product of two n-dimensional vectors needs n multiplications and n-1 addition operations. For a document collection with a total number of N documents, the time complexity in the retrieval calculation is $o(Nn)$.

In the case that the total number of document sets N=10 and the total number of keywords N=10000, search for some of them, as shown in table 5:

| plan       | Multi-keyword search | ECRS Multi-keyword search |
|------------|-----------------------|---------------------------|
| quality    | retrieval rate        | 0.040                     | 0.060                     |
|            | correct               | 0.901                     | 0.941                     |
| sales      | retrieval rate        | 0.032                     | 0.042                     |
|            | correct               | 0.844                     | 0.931                     |
| wholesale  | retrieval rate        | 0.022                     | 0.040                     |
|            | correct               | 0.741                     | 0.822                     |

According to the data comparison in table 5, it can be found that in the retrieval process, the similarity between multi-keyword vectors is calculated to retrieve, and the retrieval rate and accuracy of ECRS multi-keyword are improved. In the whole process, not only the time complexity is reduced, but also the retrieval accuracy is improved.

**6. Conclusion**

In this paper, the combination of attribute-based access control scheme, holomorphic encryption and
decryption scheme, multi-keyword retrieval scheme are applied to proposing the application between enterprise data. ECRS not only effectively prevents the theft of illegal users from accessing this part, but also ensures the security and integrity of data calculation. At the same time, multi-keyword query improves the accuracy of query results. Cipher text retrieval in ECRS is implemented in cloud storage, through employee defined query keyword conditions and dedicated server processing. Then it is converted into the standard word vector Q and sent to the cloud for retrieval. The retrieved cipher text is returned to the master control server, and cipher text analysis is realized by calculating the correlation between the attribute sets of cipher text files. Thus, it improves the accuracy of data retrieval, strengthens the security and integrity of data. Realizing the relevant departments can consult each other's department information, security and rapid realization of the data exchange between the relevant departments. However, this paper is not perfect in document update and modification, which needs further study.

Acknowledgments
This project is funded by the national natural science foundation of China (61862007, 61462009), Guangxi natural science foundation project (2018GXNSFAA138147) and Guangxi University for Nationalities postgraduate scientific research innovation project (No. Gxxzs2017114)

References
[1] Bethencourt J, Sahai A, Waters B. Cipher text policy attribute-based encryption[C]. IEEE Symposium on Security and Privacy,Los Alamitos, 2007:321-334.
[2] Rivest R, Shamir A, Dertouzos M. On data banks and privacy homomorphisms [J]. Foundation of Secure Computation,1978,7(1):169-177
[3] MICCIANCIO D. A first glimpse of cryptography's Holy Grail [J]. Communications of the ACM, 2010, 53(3): 96-96.
[4] Gentry, Craig. Fully homomorphic encryption using ideal lattices[J]. Proceedings of the Annual ACM Symposium on Theory of Computing, 2009, 9(4): 169-178.
[5] Coron, Mandal A, et al. shorter public keys[C]// Proc. of Conference on Advances in Cryptology.Springer-Verlag, 2011: 487-504.
[6] Li Wencheng, Zhao Fengyu. Research on encryption and cipher text full-text retrieval of enterprise cloud storage data[J].Small computer system, 2013(02): 529-432.
[7] Cao N, Wang C, Li M, et al. Privacy-preserving multi keyword ranked search over encrypted cloud data[C]. IEEE INFOCOM.IEEE, 2011:829-837.
[8] LYUBASHEVSHY V, PEIKERT C, REGEV O. On ideal lattice and learning with errors over rings [C]. In Eurocrypt 2010, LNCS 6110, Spring-Verlag, 2010:1-23.
[9] Zhang Jianhang, He Jian, Hu Yupu. New NTRU encryption scheme based on R-LWE[J]. Electronic technology.2012:5-15.