Encrypted Data Search Method Based on Keyword Semantic Relation Extension

Junfeng Lv
Information communication branch,
State Grid Corporation of China,
Beijing, China
E-mail: 873239023@qq.com

Yang Liu
School of Control and Computer Engineering,
North China Electric Power University,
Beijing, China
E-mail: 1558487881@qq.com

***Abstract*—The efficient retrieval of ciphertext in the cloud storage environment is a hot topic in current academic research. However, most of the existing ciphertext retrieval technologies are based on the complete matching of user query keywords, which leads to the neglect of some documents containing semantic association words. Due to the lack of general users' knowledge of relevant fields, the submitted keywords are limited, and they can not fully and accurately respond to the actual query requests of users, which leads to the problem of incomplete and inaccurate search results. To solve this problem, this paper proposes an encrypted data search method based on the semantic relationship expansion of keywords, which can effectively improve the retrieval recall rate and make the retrieval results more consistent with the user's query intention. The experimental results show that our method can return completely matching files related document and query keywords**

Keywords—Cloud storage; Semantic relationships extend; Encrypted data search;

I. INTRODUCTION

With the rapid development of the information age, the demand for enterprises and individual users for storage is increasing. To save costs and improve data access speed, more and more enterprises and individuals start to migrate data to the cloud servers. Since these data contain a large amount of sensitive information of enterprises and individuals, once these data are stored in the third-party cloud server, it will be managed and controlled by the third-party cloud server. In recent years, the frequent leakage of user data privacy makes the security of cloud server widely concerned. To protect the security of data, users usually choose to store data in the cloud server in the form of ciphertext after local encryption. Therefore, how to design an efficient and secure cloud server ciphertext retrieval becomes a valuable research topic.

In the past, the keyword search process is usually: the user first enters the keyword to query, the server will compare this keyword with the documents in the background database one by one. In the comparison process, the keyword should be matched with each word segmentation in the document [1-5]. The efficiency of this retrieval method is very low, and the retrieval accuracy is not high. In order to improve the search experience, some scholars further proposed the "approximate keyword search technology" [6-8]. The core idea of this technique is to perform fuzzy matching by calculating the character difference between keywords and document context [9-11]. Although the efficiency of the search scheme has been greatly improved, the accuracy of the search is still not very ideal. In recent years, with the development of semantic analysis, keyword semantic-based search algorithms emerge one after another. Keywords for semantic analysis, can dig up meaning similar to other extensions and of keywords, the advantage is: when the user input keywords search instead of the actual demand of large difference, the application of semantic analysis method can rapidly expand the search scope, and get a wider range of search results, so that more close to the users expect search intentions.

The basic requirement of encrypted text search technology in the cloud environment is to ensure the efficiency and practicality of search. In recent years, research focuses on ciphertext sorting retrieval technology [12-13], multi-dimensional query request retrieval technology [14-15] and secure multi-party concurrent efficient retrieval technology [16-18]. In terms of flexible, dynamic and updatable search technology, some scholars have proposed a search technology based on fuzzy keywords[19]. In the aspect of the secure proxy, some scholars have proposed a secure proxy-based on wildcard searchable encryption [20]. In addition, a variety of ciphertext search algorithms based on public-key cryptography have been developed, such as linear online matching ciphertext search algorithm[21], ciphertext search algorithm[22] for preserving encryption index, fuzzy dynamic ciphertext search algorithm [23], a new KGA-proof PEKS public-key searchable encryption [24], and ABE[25], the key strategy of monotone Boolean formula, etc.

The keyword semantic relationship library construction scheme introduced in this paper is based on the keyword word frequency analysis frequently appearing in the document. The association rule mining Apriori algorithm is used to semantically expand the search keywords, and the improved TF-IDF index calculation algorithm is used to calculate the query expansion. The correlation between the texts causes the final search process to be performed on the extended semantic relational library

II. KEYWORD SEMANTIC RELATION TO LIBRARY CONSTRUCTION

A. Keywords semantic relational library building model

As shown in Fig. 1, the system construction model of the keyword semantic relation library proposed in this paper mainly includes data owners, cloud service providers, and ordinary users. Data owners can be an enterprise or individual users, data owners in the data file **F** =

$\{F_1, F_2, F_3 \ldots F_n\}$ before outsourcing to the third-party cloud server platform, the file shall be encrypted to prevent unauthorized users from using it. When the data owner sends the private key of the file to the authorized user, the authorized user converts the key of the file to a search request by using a single generation function and sends the private key and the search request together to the cloud server. Due to the lack of understanding of relevant domain knowledge by ordinary users, the keywords submitted are very limited and they cannot make comprehensive and accurate responses to the actual query requests of users, resulting in incomplete and inaccurate retrieval results. Therefore, it is necessary to provide an encrypted data search method with the keyword semantic relation extension. When the cloud server performs an encrypted search, it organizes the collection of files found that have the same semantics as the keyword to the authorized user in a certain order.

The data owner preprocesses the document set, divides the words in each document set, filters the useless words and calculates the weight of keywords in each document. Due to the security problem of the third-party server, the keywords and their weights in the document set cannot be uploaded directly. In this paper, the keywords and their weights are encrypted by attribute-based double strategy ABE encryption algorithm and uploaded to the cloud. After receiving the ciphertext data uploaded by the data owner, the server needs to build index and semantic relational database for the data. When the authorized user sends the keyword of the query to the third-party cloud server, the cloud server will conduct semantic extension of the query keyword and search the keyword in the ciphertext document one by one according to the extended keyword set. Set the keyword weight threshold and return the document set of keywords greater than the weight threshold to the authorized user.

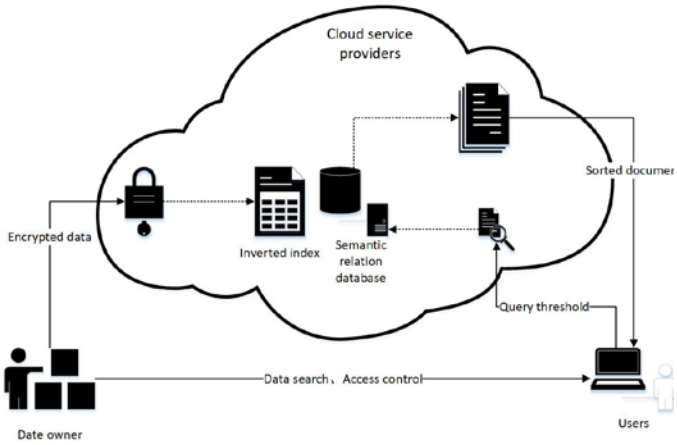

Fig. 1. Keywords semantic relational library building model.

B. Keyword Set Extension Based on Apriori Algorithms

In this paper, the Apriori algorithm is used to extend the semantics of keywords submitted by authorized users. The idea of the Apriori algorithm is to mine the frequent itemsets needed to generate association rules. This algorithm is based on the previous experience of frequent itemsets. In terms of user query keywords, each main keyword record can be composed of several common keyword items. When initializing, the set of keyword items is an itemset. Apriori algorithm mining frequent itemsets mainly in a simple level of sequential search cycle method, that is, according to the k-itemset to deduce (k+1)-itemset. The specific method is: first find out one frequent itemset, record it as L_1, then use L_1 to mine L_2, that is, two frequent itemsets, and continue to loop until no more frequent k item set can be found.

For a user's search, the frequency and distribution of keywords entered by a user are relatively concentrated. In general, it will have a high degree of relevance with the current search topic. And every user has their habits. Generally speaking, the distribution of keywords input will not be too wide. So we can use the Apriori algorithm to search and mine the frequent set of keywords that each user has reached a certain frequency. The keyword set expansion algorithm based on Apriori algorithm is described as follows:

Algorithm: Keyword set extension

Input:

1、the set of keywords entered by the user is $KEY = \{k_1, k_2, \ldots, k_n\}$

2、current keyword dictionary table D

1) Find a frequent set with min_sup as the minimum support L_1。

$L_1 = \text{find\_1\_itemset}(D, \min\_sup);$

2) Generating frequent k sets from frequent (k-1) sets. $\text{for}(k = 2; L_{k-1}; k++)\{$

$L_k = \text{apriori}(L_{k-1}, \text{min\_sup});\}$

3) Find the number of items that contain the most frequent itemsets. $m = \max\{L_k\};$

4) if $(n >= m)$ then

$\text{EXTEND\_KEY} = \text{empty and goto 8;}$

5) Find the most frequent itemset.

$temp = \text{max\_frequent}(L_k)$,while n<k<m;

6) $L' = \cup L_i, L_i \in temp;$

7) if (L') is empty then $\text{EXTEND\_KEY} = \text{empty}$

else $\text{EXTEND\_KEY} = L' - KEY$

8) Return to EXTEND_KEY.

Output: extended keyword set EXTEND_KEY

The core idea of this algorithm is mainly embodied in the following: specific users usually have certain search habits for query requests on related topics. In this way, when a user enters a new set of keywords, the previous history is used to find the most frequent set with the highest frequency of the currently entered keyword item set. Then, you can treat the words in the frequent set that do not appear in the keyword entered by the user as not matching the current search direction of the user. Compared with the traditional keyword similarity calculation, the time complexity of keyword recommendation in the Apriori algorithm is O(n) ~ O(n2), which greatly improves the efficiency.

C. Computation of Comprehensive Relevance between Documents and Extended Queries

There are many calculation indexes of comprehensive correlation of document extended query. TF-IDF algorithm

is used to calculate the keyword weight in the document set, and keyword weight is used as the basis of query expansion. In this paper, based on the improved TF-IDF index calculation algorithm, the original keyword frequency in the document to be expanded is set as TF, which can be used to reflect the keyword weight of the document. IDF index is generally set as the inverse of the frequency of the keyword in the overall document set [14], which can be used to reflect the weight of the keyword in all documents. The larger the index value is, the more it can reflect the contribution of the keyword to the keywords in the overall document set. Then the formula for the comprehensive correlation of the keyword w in document F_i is as follows:

$$S(w, F_i) = \frac{(1+\ln f_{i,w}) \ln\frac{1+n}{f_{i,w}})}{1+|F_i|} \quad (1)$$

Where $f_{i,w}$ is the number of keyword w, and $|F_i|$ is the total number of non-blank characters contained in document F_i.

III. Experiment

A. Experimental environment

The experiment is programmed in C++ high-level language and tested with the RFC standard document set [26]. The total number of documents selected is 5000. The specific configuration of the experimental environment is shown in TABLE I.

TABLE I. Experimental Environment Configuration

Hardware and software	*Specific configuration*
CPU	AMD R5,3.0 GHz
Memory	2.5 GB
OS	Windows 7,64 bit

B. Generate document initialization data

It takes a certain amount of storage space and time to generate document initialization data. Specifically, in addition to the semantic generation of keywords in each document dataset, we also need to calculate the semantic association values and compare them according to the degree of correlation of keywords encrypted. The size of initialization data depends on the frequency of keywords. For a global search of overall documents, query efficiency is closely related to the total number of documents. Table II lists the average resources needed to initialize the data for the document. The purpose of using the average value as the index is to eliminate the performance uncertainty caused by different experimental samples[16].

TABLE II. Document Initialization Data to Create the Average of Required

Number of files	*Space usage*	*Generation time*
1000	185B	0.27s
2000	198B	0.31s
3000	205B	0.35s

C. Generating Index and Constructing Semantic Relational Database

The document initialization data set is dynamically filtered, and the keywords with a high correlation degree are selected as the keywords of the index. Fig. 2 shows that the time of index creation is linear with the size of the document initialization dataset.

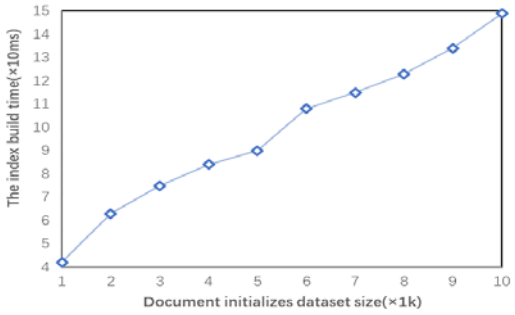

Fig. 2. The relationship between index creation time and document initialization dataset size.

The association rule mining algorithm [17] is used to initialize the set of semantic relations to be empty, and then parallel expansion is carried out for each keyword item. After the frequent itemset of keywords is formed, it is mapped with each item in the document initialization data and the correlation degree is calculated. Finally, according to the degree of correlation, the semantic relational library is generated. Fig. 3 shows that the creation time of the semantic relational library is roughly proportional to the size of the document's initialization dataset.

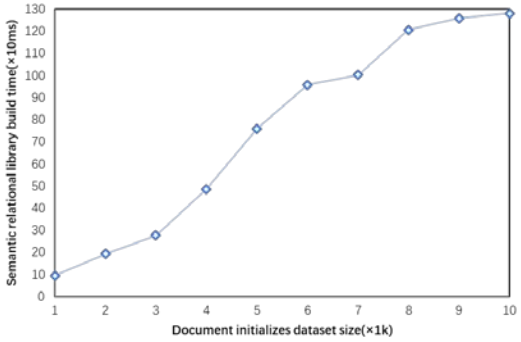

Fig. 3. The relationship between the creation time of the semantic relational database and the size of the document initialization dataset.

D. Analysis of experimental results

The main performance indexes in retrieval are index generation and resource consumption in semantic relational library construction. The experiment shows that the time of index creation is linearly related to the size of document initialization data set, and the time of semantic relation library is also directly proportional to the size of document initialization data set, which indicates that the key factor affecting query efficiency is the number of query keywords.

IV. Conclusion

This paper first introduces the shortcomings of traditional retrieval methods and the development and research results

of privacy protection ciphertext data retrieval technology and then proposes a ciphertext data retrieval model based on semantic relationship extension. Then, the key extension scheme based on the Apriori algorithm and the scheme using TF-IDF index to calculate the comprehensive correlation between document and extended query is introduced in detail. Experimental results show that our solution can return not only exactly matched files, but also files that are completely relevant to the query keywords in the solution, and construct corresponding file metadata for each file.

ACKNOWLEDGMENT

Fund Project: Supported by the science and technology project "research and application of key technologies of open source software security monitoring" of State Grid Corporation of China(SGFJXT00YJJS1800074);

REFERENCES

[1] Shekarpour. S, Hoffner. K, Lehmann. J and Auer. S, "Keyword query expansion on linked data using linguistic and semantic features" in 2013 IEEE Seventh International Conference on Semantic Computing, 2013.

[2] Yadav. C.S, Sharan. A, Joshi M.L, "Semantic graph based approach for text mining" in Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on, 2014.

[3] Qian Chen, Zengru Jiang and Jinqiang Bian, "Chinese keyword extraction using semantically weighted network" in Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2014 Sixth International Conference on, 2014.

[4] Sheeba, J.I and Vivekanandan. K, "Low frequency keyword extraction with sentiment classification and cyberbully detection using fuzzy logic technique" in 2013 IEEE International Conference on Computational Intelligence and Computing Research, 2013.

[5] Pakgohar, Alireza, Khalili, Mohadeseh, "A probabilistic relational model for keyword extraction" in Statistics in Science, Business, and Engineering (ICSSBE), 2012 International Conference on, 2012.

[6] Zhipeng Chen, Zhiyang He, Ping Lv and Ji Wu, "Improving keyword search by query expansion in a probabilistic framework" in Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on, 2014.

[7] Qin Liu, Guojun Wang and Jie Wu, "Secure and privacy preserving keyword searching for cloud storage services", Journal of Network and Computer Applications., 2011.

[8] Ren, Kui, Wang, Cong , Wang and Qian, "Security challenges for the public cloud", IEEE Internet Computing., 2012.

[9] Mihir Bellare, Joe Kilian and Phillip Rogaway, "The security of the cipher block chaining message authentication code", Journal of Computer and System Sciences,. 2000.

[10] Wang, Cong,Cao, Ning,Ren, Kui,Lou and Wenjing, "Enabling secure and efficient ranked keyword search over outsourced cloud data", IEEE Transactions on Parallel and Distributed Systems,. 2012.

[11] Liming Fang, Willy Susilo, Chunpeng Ge and Jiandong Wang, "Chosen-ciphertext secure anonymous conditional proxy re-encryption with keyword search", Theoretical Computer Science,. 2012.

[12] KAMARA S, PAPAMANTHOU C and ROEDER T, "Dynamic searchable symmetric encryption" in Proceedings of the 2012 ACM conferenceon Computer and Communications Security(CCS' 12), 2012.

[13] Bruno M. Fonseca, Paulo B. Golgher, Edleno S. De Moura, Bruno Pssas and Nivio Ziviani, "Discovering search engine related queries using association rules", Journal of Web Engineering,. 2003.

[14] Ming Li, Shucheng Yu, Yao Zheng, Kui Ren and Wenjing Lou, "Scalable and secure sharing of personal health records in cloud computing using attribute-based encryption", IEEE Transactions on Parallel and Distributed Systems,. 2013.

[15] CAO N, WANG C, LI M, et al, "Privacy-preserving multi-keyword ranked search over encrypted cloud data", IEEE Transactions on Parallel and Distributed System., 2014.

[16] SUN W, WANG B, CAO N, "Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking" in proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security, 2013.

[17] Hohenberger, S. Waters and B. Source, "Attribute-based encryption with fast decryption" in Public-Key Cryptography - PKC 2013. 16th International Conference on Practice and Theory in Public-Key Cryptography, Proceedings, pp 162-79, 2013.

[18] Golle, P. Staddon, J.Waters and B. Source, "Secure conjunctive keyword search over encrypted data. "Applied Cryptography and Network Security in Second International Conference, ACNS, 2004.

[19] Kamara, Seny, Papamanthou, Charalampos Source, "Parallel and dynamic searchable symmetric encryption" in Lecture Notes in Computer Science, v 7859 LNCS, p 258-274, 2013, Financial Cryptography and Data Security - 17th International Conference, FC 2013.

[20] Shen-Ming Chung, Ming-Der Shieh and Tzi-Cker Chiueh, "A security proxy to cloud storage backends based on an efficient wildcard searchable encryption" in 2018 IEEE 8th International Symposium on Cloud and Service Computing (SC2), 2018.

[21] Cash, D., Jarecki, S., Jutla, C., Krawczyk, H. et al. "Highly-scalable searchable symmetric encryption with support for Boolean queries" in Advances in Cryptology - CRYPTO 2013. 33rd Annual Cryptology Conference. Proceedings: LNCS 8042, p 353-73, 2013.

[22] CASH D, JARECKI S and JUTLA C, "Dynamic universal accumulators for DDH groups and their application to attribute based anonymous credential systems" in Topics in Cryptology-ct-rsa, the Cryptographers Track at the Rsa Conference ,2009.

[23] Jun Shao, Zhenfu Cao, Xiaohui Liang and Huang Lin, "Proxy re-encryption with keyword search", Information Sciences,. 2010.

[24] Takanori Saito and Toru Nakanishi, "Designated-senders public-key searchable encryption secure against keyword guessing attacks" in 2017 Fifth International Symposium on Computing and Networking (CANDAR), 2017.

[25] Jun'ichiro Hayata, Masahito Ishizaka, Yusuke Sakai, Goichiro Hanaoka and Kanta Matsuura, "Generic construction of adaptively secure anonymous key-policy attribute-based encryption from public-key searchable encryption" in 2018 International Symposium on Information Theory and Its Applications (ISITA), 2018.

[26] Kurt, M. and Yerlikaya, T, "A new modified cryptosystem based on Menezes Vanstone elliptic curve cryptography algorithm that uses characters' hexadecimal values" in 2013 International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE 2013), p 449-53, 201