Survey on Semantic-Aware Searching Over Encrypted Data on Cloud Systems

[1]Komal K. Sharma, [2]Prof. Mrunalinee Patole

Department of Computer Engineering ,RMDSSOE, Pune, Maharashtra, India

Abstract: Storing and retrieving such a large amount of data consumes lot of time as data in the cloud needs to be always stored in encrypted format while storing and needs to be decrypted while searching. There are a number of propositions for executing queries over encrypted data. This implements the client to encrypt data before outsourcing it to the cloud in a database scheme. To avoid this massive consumption of time, data searching speed can be increased by directly searching over encrypted data in the cloud. There are many methods used for searching the encrypted data over cloud. In keyword-based search schemes ignore the semantic representation information of users retrieval, and cannot completely meet with users search intention. In this paper, propose ECSED, a novel semantic search scheme based on the concept hierarchy and the semantic relationship between concepts in the encrypted datasets. ECSED uses two cloud servers. One cloud server is used to store the outsourced datasets and return the ranked results to data users. The other cloud server is used to compute the similarity scores between the documents and the query and send the scores to the first server. To further improve the search efficiency, system utilizes a tree-based index structure to organize all the document index vectors. Then employ the multikeyword ranked search over encrypted cloud data as our basic frame to propose two secure schemes.

Keywords: Searchable encryption, cloud computing, smart semantic search, concept hierarchy

I. INTRODUCTION

Cloud computing refers to accessing software and storing data in the cloud of the internet. It is a model for enabling convenient, on demand network access to a shared pool of configurable and reliable computing resources that can be rapidly provisioned and released with service provider interaction. The security of outsourced data cannot be guaranteed, as the Cloud Service Provider (CSP) possesses whole control of the data. So, it is necessary to encrypt data before outsourcing them into cloud to protect the privacy of sensitive data . The idea of proposed system comes from many researchers have proposed a series of efficient search schemes over encrypted cloud data. All the existing searchable encryption schemes, which consider keywords as the document feature, do not take the semantic relations between words into consideration. The semantic relations between words are diverse [8], such as synonymy and domain correlation. Considering the potentially huge amount of outsourced data documents in the cloud, the search accuracy and search efficiency are influenced negatively if the semantic relations between words are not handled well. In this paper, proposes an efficient searchable encrypted scheme based on concept hierarchy supporting semantic search with two cloud servers. A concept hierarchy tree is constructed based on domain concepts related knowledge of the outsourced dataset. The concept hierarchy is extended to include more semantic relations between concepts. With the help of extended concept hierarchy, document features are extracted more precisely and search terms are well extended based on the semantic relations between concepts.

II. LITERATURE SURVEY

In recent years, many researchers have proposed a series of efficient search schemes over encrypted cloud data. Research paper, ‘Semantic-aware Searching over Encrypted Data for Cloud Computing’ published by Zhangjie Fu, [1] ,in this to address the problem of semantic retrieval, author propose effective schemes based on concept hierarchy. To improve accuracy, author extend the concept hierarchy to expand the search conditions. Paper, ‘Towards Efficient Content-aware Search over Encrypted Outsourced Data in Cloud’ published by Zhangjie Fu, [2] in this paper, author uses an new semantic search scheme based on the concept hierarchy and the semantic relationship in concepts in the encrypted datasets. More specifically, our scheme first indexes the documents and builds trapdoor based on the concept hierarchy. Paper, ‘Dual-Server Public-Key Encryption with Keyword Search for Secure Cloud Storage’ published by R. Chen, [3] the searchable encryption which allows the user to retrieve the encrypted documents that contain the user-specified keywords, where given the keyword trapdoor, the server can find the data required by the user without decryption. Author investigate the security of a well-known cryptographic primitive, namely Public Key Encryption with Keyword Search (PEKS) which is very useful in many applications of cloud storage. Paper, ‘Identity-based Encryption with Outsourced Revocation in Cloud Computing’ published by Jin Li, [4] in this, Identity-Based Encryption (IBE) which simplifies the public key and certificate management at Public Key Infrastructure (PKI) is an another relevant to public key encryption. Paper, 'Privacy-Preserving Smart Semantic Search Based on Conceptual Graphs Over Encrypted Outsourced Data' published by Zhangjie Fu[5], in this, Considering various semantic representation tools, author select Conceptual Graphs as our semantic bearer because of its great ability of expression and extension. To improve the efficiency of retrieval, author uses Tregex simplify the key sentence and make it more generalizable. Here transfer of CG into its linear form with some alteration which makes quantitative calculation on CG and fuzzy retrieval in semantic level possible.

Paper, 'Secure kNN Computation on Encrypted Databases' published by, W. K. Wong [6] in this, author discuss the general problem of secure computation on an encrypted database and propose a SCONEDB (Secure Computation ON an Encrypted DataBase) model, which captures the execution and security requirements. Author focus on the problem of k-nearest neighbor (kNN) on encrypted datasets. Paper, 'Building and Applying a Concept Hierarchy Representation of a User Profile' published by, Nikolaos Nanas [7] in this, author creates method for the construction of a concept hierarchy that takes three basic dimensions of term dependence. Paper, 'WordNet: A Lexical Database for English' published by, George A. Miller[8] in this, WordNet provides a more effective combination of traditional lexicographic information and modern computing. WordNet is an online lexical database design to use under program control. Paper, 'Fuzzy Keyword Search over Encrypted Data in Cloud Computing' published by, Jin Li [9] in this, author exploit edit distance to quantify keywords similarity and develop leading technique while constructing fuzzy keyword sets, which reduces the storage. Paper, 'Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data' published by, Ning Cao [10] in this, author define and solve the challenging problem of privacy-preserving multi-keyword ranked search over encrypted data in cloud computing (MRSE).

III. SYSTEM MODEL

Author uses two cloud servers to serach, in system model there are four entities as shown in figure 1: the data owner, the data user, the cloud server A, the cloud server B.

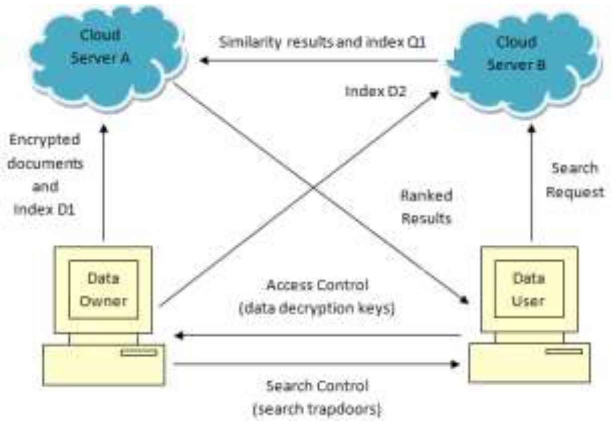
Cloud Server A
Similarity results and index Q1
Cloud Server B
Index D2
Encrypted documents and Index D1
Ranked Results
Search Request
Data Owner
Access Control (data decryption keys)
Data User
Search Control (search trapdoors)

Fig 1: System Model

Data owner: The data owner encrypts the data held locally and uploads it to the cloud server.
A concept hierarchy is constructed based on the domain concepts related knowledge of the dataset and two index vectors (D1,D2) for each document of the dataset are generated based on the key concepts of the document and the concept hierarchy. Then, the searchable index which is constructed with all the index vectors is sent to the cloud A.

Data users: The authorized data user makes a search request. Then, the trapdoors which related to the keywords are generated. At last, the data user sends the trapdoors to the cloud B.

Cloud Server A: The cloud server A has two functions. One is storing the outsourced dataset. The other one ranks the results from the cloud B and returns the certain encrypted documents that satisfy the search criterion to data users.
Cloud Server B: The cloud server B is used to compute the similarity scores between documents vector and trapdoors vector when it receives the trapdoor. After computing, the cloud B submits these results to the cloud A.

3.1 CONCEPT HIERARCHY

A concept hierarchy is an organized concept set using hierarchical method. In the hierarchy, the concepts at lower levels contain more specific meanings than those at higher levels. At first, we generate the concept hierarchy based on the domain information of the outsourced dataset. And then we deal with the dataset to extend the concept hierarchy. Concept hierarchy can be created by its own, based on the number of distinct values per attribute in the known attribute set. The attribute with the utmost specific values is placed at the lowest level of the hierarchy.

- Auto generate the attribute ordering based upon observation that attribute defining a high level concept has a smaller # of distinct values than an attribute defining a lower level concept
- Example : Fig 2 shows concept hierarchy for college employees.

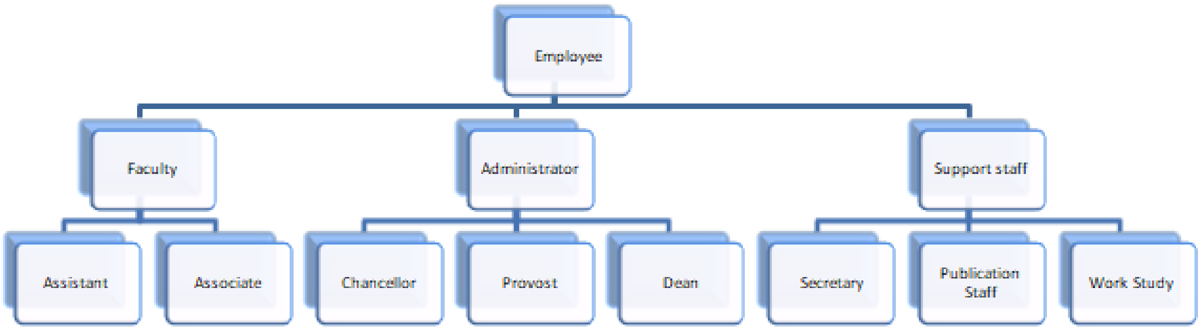

Fig 2. A concept hierarchy for college employees

The domain concept hierarchy can be obtained by some existing tool, such as WordNet [8], The similarity between two concepts is calculated based on the distance of them in the concept hierarchy. A semantic relation in the concept hierarchy is shown in Table 1.

Semantic relation type	Example
Synonym	web-network
Hyponymy	fruit-apple
meronym / holonym	forest-tree
host-attribute	people-name
attribute-value	color-red

Table 1 .A semantic relation in the concept hierarchy

3.2 SECURE SEARCH SCHEME

1) **Generating Document Index Vector:** "Attribute-value" relation in the hierarchy,two index vectors should be generated for each document in the dataset, one vector is used to match concepts in the search request and another one is used to determine whether the value for an attribute is satisfied with the search request. For a document F, we denote its two index vectors by D1 and D2. Each dimension of D1, denoted by D1[i], corresponds to a node (stores concept ci) in the hierarchy. If F contains the concept ci, then D1[i] = 1,otherwise D1[i] = 0. Similarity, each dimension of D2, denoted by D2[i], corresponds to a node (stores concept ci) in the hierarchy.
2) **Generating Trapdoor:** For a search request containing several concepts, two n-dimension vectors are also generated, one is used to store the information about concepts in the search request and another one is used to store the search restriction on attribute.Given the index vectors of a document F and the search trapdoor of a search request Q, the search procedure is conducted as follows. Firstly, the procedure checks whether the document satisfies search restrictions included in search request using vectors D2 and Q2. Secondly, if D2 satisfies Q2, then the procedure computes D1 .Q1 to obtain the similarity score of the document to the search request, where the value of D1 _.Q1 indicates the number of matched concepts between F and Q. At last, all the related documents are sorted based on their similarity scores and the top-k related documents are returned to the user, where k is a parameter received from the user.
3) ECSED-**1 Scheme: Secure Scheme in Known Ciphertext Model:** In this paper, we replace the previous framework ASPE[2] with the framework MRSE to improve the efficiency and security. Here use MRSE [10] as our basic framework. To improve the efficiency and security, then split the retrieval process into two parts and carry out them in two servers, respectively. Meanwhile, we extend the dimension of the vectors to (n + 2) for reducing the possibility that the cloud servers can infer the relationship between the trapdoor and the index. Thus, we propose the secure scheme ECSED-1 under the known ciphertext model. The specific algorithm are as follows:
   - **Setup:** security parameter as input, and then algorithm generates a public parameter P.
   - **SKeyGen**: The data owner randomly generates a secret key SK, which is in the form of a 3-tuple as {S,M1,M2}.
   - **BuildIndex**: For each document F in the dataset, the data owner generates two index vector D1 and D2.
   - **Trapdoor**: For a search request Q, the data user generates two vectors Q1 and Q2 . Use MRSE to encrypt the vector Q1.
   - **BTest**: This process has been done under the cloud server B. In the BTest, the procedure checks whether a document F should be returned in two steps : 1) for each pair (Q2a, Q2b)is used to determine whether D2 meets the restriction condition; 2) if D2 satisfies all restriction conditions, then author can get the similarity between the query and document F with the given t for the query vector and the given the data vector.
   - **ATest**: Execute the ATest under the cloud server A. The main function of this algorithm is to sort the results which returned from BTest, and then return the first k files which meet the user requirements to data users.
4) ECSED-**2 Scheme: Secure Scheme in Known Background Model:** Compared with the conference version[2], this paper make new schemes for new threat models. For the known

background model, the security of the above scheme is not high enough because the cloud server understand some information of the background relationships between the trapdoors and the specific keywords, it is possible to infer the specific keyword by the hidden information of the trapdoors. As a instance, the cloud servers can guess some high frequency keywords, through the known background information and the documents frequency. Therefore, we propose a more secure solution ECSED-2 to resist the known background attack. These algorithm is as same as previous one.

3.2 DISCUSSION ON RESULT:

- Encryption scheme generate a search index based on the keyword dictionary which is extracted from outsourced datasets, with trapdoor generated in the search stage, the server can search the searchable index and return related documents.
- Semantic search becomes more important, as traditional keyword based search scheme cannot exploit the hidden meaning of terms.
- Concept hierarchy a semantic search tool used for organizing concepts, constructed to indicate the relationship between concepts.

IV. CONCLUSIONS:

In this survey, to address the problem of semantic retrieval, propose effective schemes based on concept hierarchy. Solution uses two cloud servers for encrypted retrieval and make contributions both on search accuracy and efficiency. To improve accuracy, we extend the concept hierarchy to expand the search conditions. In addition, a tree-based index structure is constructed to organize all the document index vectors, which are built based on the concept hierarchy for the aspect of search efficiency. The security analysis shows that the proposed scheme is secure in the threat models.

REFERENCES:

[1] Zhangjie Fu, Lili Xia, Xingming Sun, Alex X. Liu, Guowu Xie, “Semantic-aware Searching over Encrypted Data for Cloud Computing”, IEEE Transactions on Information Forensics and Security

[2] Z. Fu, X. Sun, S. Ji, and G. Xie, ”Towards efficient content-aware search over encrypted outsourced data in cloud,” Proc. of IEEE INFOCOM 2016, pp.1-9,2016.

[3] R. Chen, Y. Mu, G. Yang, F. Guo, and X. Wang. ”Dual-server public-key encryption with keyword search for secure cloud storage,” IEEE Transactions on Information Forensics and Security, vol.11,no.4, pp.789-798,2017.

[4] J. Li, J. Li, and X. Chen, ”Identity-based encryption with outsourced revocation in cloud computing,” Computers, IEEE Transactions on, vol.64,no.2,pp.425-437,2015.

[5] Z. Fu, F. Huang, K. Ren, J. Weng, and C. Wang. ”Privacy preserving smart semantic search based on conceptual graphs over encrypted outsourced data,” IEEE Transactions on Information Forensics and Security, vol.12,no.8, pp.1874-1884,2017.

[6] W. K. Wong, D. W. Cheung, B. Kao, and N. Mamoulis, “Secure knn computation on encrypted databases,” in Proc. of SIGMOD, 2009, pp. 139–152

[7] N. Nanas, V. Uren, and A. D. Roeck, “Building and applying a concept hierarchy representation of a user profile,” in Proc. Of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 2003.

[8] G. A. Miller, “WordNet: a lexical database for English,” Communications of the ACM, vol.38, issue 11, pp. 39–41, 1995.

[9] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, “Fuzzy keyword search over encrypted data in cloud computing,” in Proc. of IEEE INFOCOM’10 Mini-Conference, San Diego, CA, USA, March 2010, pp. 1–5.

[10] N. Cao, C. Wang, and M. Li, ”Privacy-preserving multi-keyword ranked search over encrypted cloud data,” Parallel and Distributed Systems,IEEE Transactions on, vol.25,no.1, pp.222-233,2014.