International Journal of Emerging Engineering Research and Technology
Volume 2, Issue 7, October 2014, PP 215-219
ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)

Semantic Search Supporting Similarity Ranking Over Encrypted Private Cloud Data

Nagesh Jadhav[1], Jyoti Nikam[2], Sayli Bahekar[3]

[1]Department of CSE, MITCOE, Pune, India
[2]Department of CSE, MITCOE, Pune, India
[3]Department of CSE, MITCOE, Pune, India

Abstract: *With the appealing features of cloud computing, cloud becomes an important infrastructure of enterprise IT. A large amount of data is being outsourced to the cloud. Before outsourcing the data is being encrypted. Encryption makes simple but important functionalities like search operations over cloud data difficult. The traditional and efficient plaintext keyword search technique has no effect on encrypted data. The existing searchable encryption schemes support only exact keyword search, not support semantic based search. Hence we propose a search scheme wherein the semantic relationship and synonym of the query keyword are considered with the help of data structures like Semantic Relationship Library (SRL), Inverted Index. The result files are displayed in order according to total relevance score.*

Keywords: *Semantic, Cloud Data, Semantic Relationship Library (SRL), Inverted Index, Relevance Score.*

1. Introduction

Today’s world is a data intensive world. Cloud Computing is a powerful tool which enables customer to manage large scale data with on-demand applications and services in cost efficient way. Sensitive and confidential data is put on cloud. This data is encrypted first to provide privacy to the data. It is very difficult to perform data utilization operations on this particular encrypted set of data. This data is also shared with multiple users who may want to access file of a particular interest. The most common way to search file is using keyword search. But plaintext keyword search becomes useless over encrypted data[1].

Searchable encryption techniques have been developed in recent years to perform search over secure outsourced data. But they only support fuzzy keyword or exact keyword search and ignore the semantic relationship of the keyword, thus many files are omitted [2]. These search schemes send the results considering only the absence or presence of the keyword and result ranking still remains out of picture[4].

In this paper, we propose a search scheme which supports secure semantic based search and similarity ranking. In the proposed scheme, a metadata set is built for every file. This metadata set is encrypted and then uploaded to the cloud. The cloud server manipulates the received data, performs necessary operations and builds the Semantic Relationship Library (SRL) and inverted index. Upon receiving the query keyword from the user, the cloud expands it and finds the semantically related words using SRL and inverted index [3]. Then the exact keyword and semantically related words are used to retrieve files. The search results are then displayed according to the total relevance score.

2. Related Work

Song.et.al. proposed the first Searchable encryption scheme, where a two layered encryption architecture was used [5]. Coa.et.al and Yang.et.al proposed a scheme for multi-keyword ranked search where “Inner product similarity” is used for result ranking [6][7]. Emil.et.al proposed a structure in hierarchical form which could achieve secure and effective dynamic updating[8]. Whereas, Boneh.et.al proposed a key in public key setting for public key based searchable encryption scheme.

The above schemes only supported exact keyword search. In order to enhance the search features, new models supporting fuzzy keyword were proposed. In fuzzy keyword search, Liet.et.al and Wang.et.al considered distance similarity metric of keywords [10][11]. The size of fuzzy keyword set was further reduced by Liu who proposed “Dictionary based fuzzy set”[9]. But all the above schemes support similarity based only on the structure of the keyword and the semantic relation of the keyword still remains unaddressed.

3. Proposed Scheme

3.1. System Model

The proposed system model consists of three different entities: 1.Data Owner, 2.Authorised Data User, 3.Private Cloud Server as shown in figure 1.

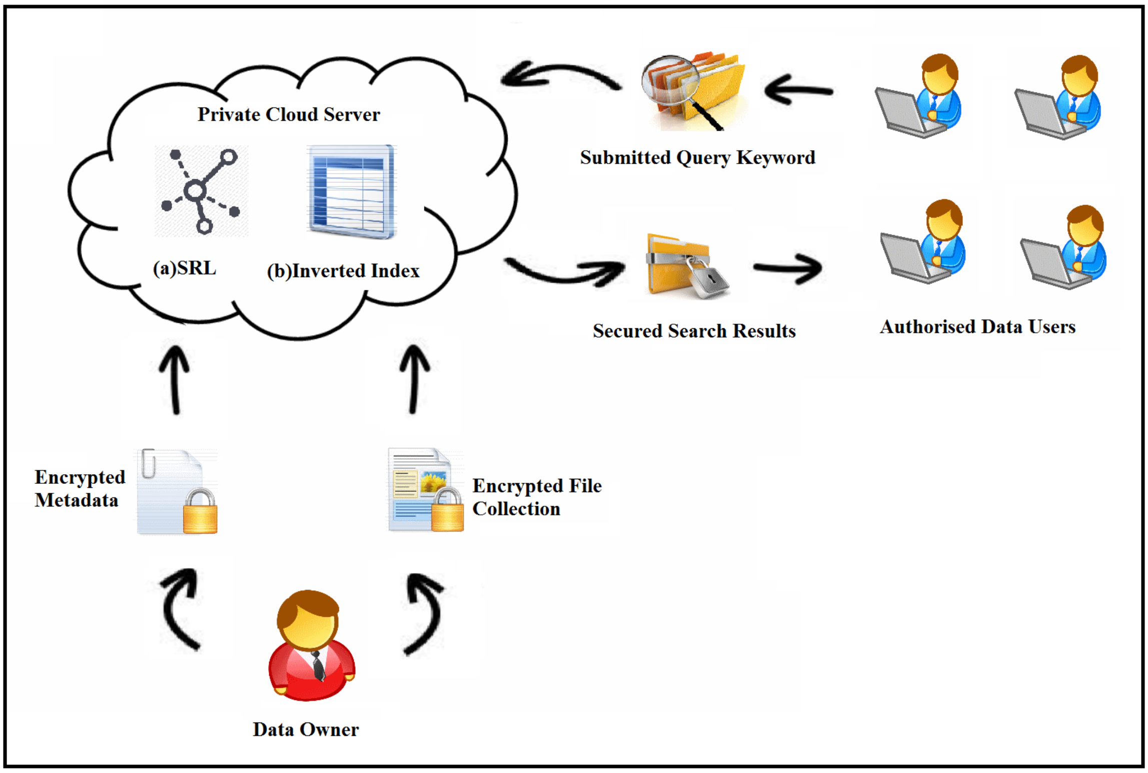

Fig1. *System Architecture for semantic expansion based search over encrypted private cloud data*

Data Owner uploads the files in encrypted format to the cloud server. A corresponding metadata is build for every file and the metadata is also uploaded to the cloud in encrypted form. The encryption algorithm which is used is asymmetric encryption algorithm, e.g. RSA.

Data User submits his interested keyword in the form of trapdoor. Trapdoor is the encrypted form of the query keyword.

The cloud server initially constructs the inverted index and builds SRL. Upon receiving the trapdoor from the authorised user, the cloud server first expands the query keyword using SRL. Then the cloud server searches the index and returns matched files to the user in descending order of relevance score. The user then decrypts the received files and gets the required data.

3.2. Notations

F – The plaintext file collection owned by data owner. Set of n no. of files are shown as, F= $\{F_1, F_2, \cdots, F_n\}$.

E – Encrypted file collection, stored in cloud server by the owner. Denoted as E= $\{e_1, e_2, \cdots, e_n\}$.

Id(F_i) – Identifier of file F_i to uniquely identify the files.

K – The keyword dictionary extracted from the files, denoted as a set of p keywords. K= $\{k_1, k_2, \cdot\cdot, k_p\}$.

M – The metadata set in encrypted form, denoted as a set of n file metadata M= $\{M(F_i)\}$, i=1,2,..,n.

I – The inverted index including a set of m lists. I=$\{I(K_m)\}$, i= 1,2,..m.

T_k – The trapdoor generated for a query keyword k by a user.

S_k– The semantically expanded keyword set of k, it is a subset of K, denoted as $S_k = \{k_1', k_2', \cdots\}$.

3.3. Primary Function

3.3.1. Construction of SRL

The basic functionality of semantic query expansion is to find out semantic relationship between the keywords. There are readily available knowledge models for this purpose, e.g. EuroWordNet, WordNet. There are some another technologies supporting dynamic construction of semantic relationship from the document collection. Such as, mutual information model and term clustering. Mutual information model is used in this paper to find the relationship between two keywords. The mutual information I(x,y) can be defined as

$$I(x,y)=\log_2\frac{P(x,y)}{p(x)p(y)} \qquad (1)$$

Here, P(x,y) is the probability of occurrence of x and y together, p(x) and p(y) are the probabilities of individual occurrence of x and y in the file collection respectively.

The value of mutual information obtained by equation (1) is then normalised into a relationship value in interval [0, 1]. A weighted graph structure is then constructed which depicts the semantic relationship library.

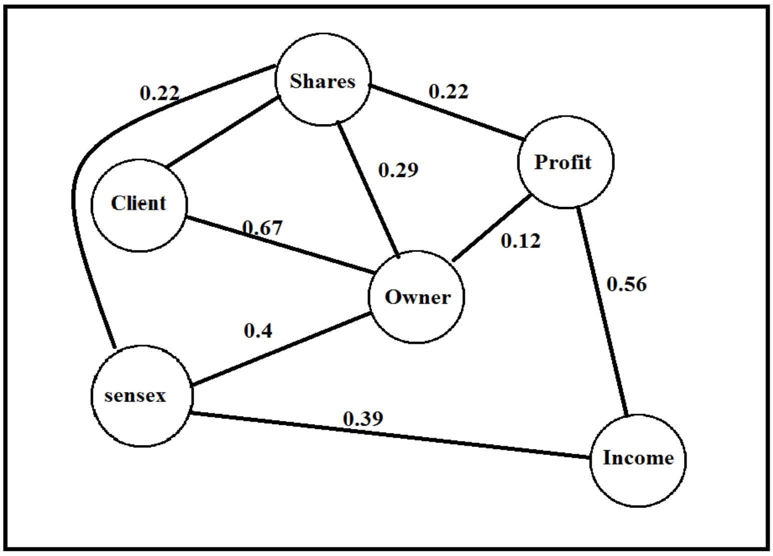

Fig2. *Example of Semantic Relationship Library*

3.3.2. Construction of Inverted Index

Inverted index is a structure showing association between files and the different keywords in the files. Relevance score is generated for each file for the purpose of ranking. The calculation of relevance score is shown in the next part.

Inverted index is generally a table structure for each keyword as shown in the table 1.

Table1. *Inverted Index*

Keyword	k_i				
File Identifier	Id(F_{i1})	Id(F_{i2})	Id(F_{i3})	…	Id(F_{in})
Relevance Score	RS_{i1}	RS_{i2}	RS_{i3}	…	RS_{in}

Here, in the above table, Rs_{in} denotes the relevance score of file F_{in} corresponding to keyword $k_{i.}$

3.3.3. Calculation of Relevance Score

The most common method used to calculate the relevance score is the TF*IDF . Here TF indicates the term frequency i.e. the no of occurrence of the keyword in a particular file and IDF indicates the importance of the keyword in the whole collection of files i.e. its occurrence in the *complete* set of files. The equation used to evaluate the relevance score for a single keyword is as follows:

$$\text{Score}(k,F_i) = \frac{1}{|Fi|} *(1+ \ln f_{i,k})* \ln(1+\frac{n}{fk}) \tag{2}$$

In the above equation, $|Fi|$ is the length of the file F_i ; $f_{i,k}$ denotes the term frequency of the word k in file F_i ; f_k denotes the no. of files containing the word k.

In our proposed scheme, the query keyword is first expanded, after this the relevance score for the expanded word is calculated. Then the total relevance score of submitted query keyword is calculated by adding relevance scores of individual query keyword and the expanded words, using equation (3).

$$\text{TotalRS}(k,F_i) = \text{Score}_k + \sum_{\forall ki' \in Sk} Score_{wi}' * R_i \tag{3}$$

In the above equation, Score_k denotes relevance score of the submitted query and Score_{wi}' represents relevance score of semantically related keyword.

3.3.4. Structure of File Metadata

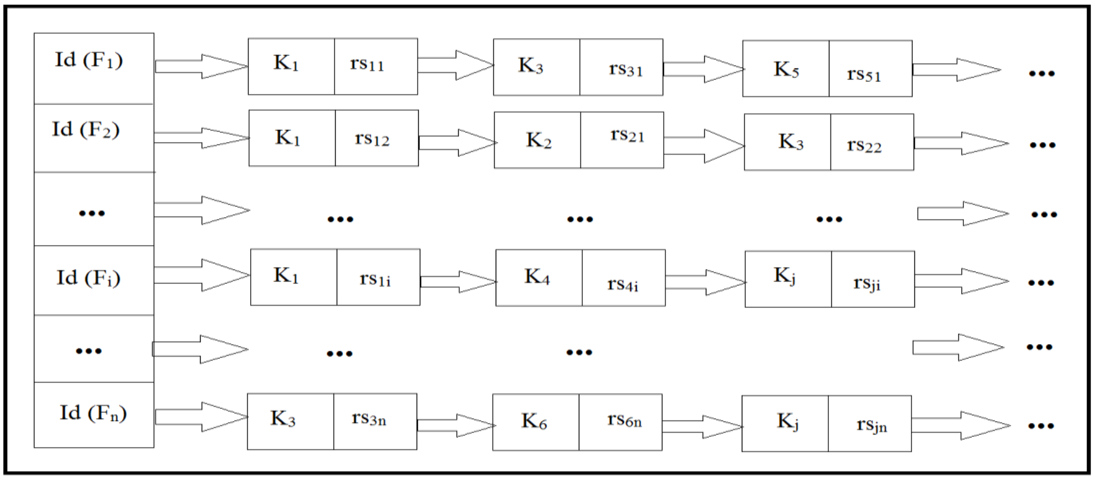

Fig3. *Example of File Metadata*

The structure of file metadata as shown in the figure 3 is an adjacency list. It consists of file identifier, keyword and its corresponding relevance score.

3.4. Semantic Expansion Based Search Scheme

The entire scheme is divided into two phases, Initialization Phase and Search Phase. These two phases consists of different algorithms.

3.4.1. Initialization Phase

It is a first phase of the scheme wherein major work of initialization is performed. The different algorithms are used in this phase are as follows:

(a) Key Generation: In this step initialization and sharing of private and public keys between the data owner and authorised data user takes place. The encryption algorithm used is an asymmetric algorithm. E.g. RSA. The file collection is then encrypted using the private keys by the data owner.

(b) Building Encrypted Metadata: From the encrypted file set, metadata is constructed. The metadata is also encrypted and outsourced to the cloud server by the data owner.

(c) SRL construction: The cloud server starts the construction of the SRL once it receives the secure metadata from the data owner. The technique of building SRL is already discussed in 3.3.1.

(d) Build Inverted Index: The cloud server extracts encrypted keywords from the metadata and generates posting list for each keyword as discussed in 3.3.2.

3.4.2. Search Phase

In this phase, the actual task of searching is performed. The authorised data user comes into picture in this phase.

(a) Query Trapdoor Generation: The user generates the encrypted version of the query keyword using the same encryption algorithm used in Initialization phase.

(b) Data retrieval: On receiving the query trapdoor, the cloud server firstly expands the query keyword and finds the words which are semantically related to the query keyword. It then searches the expanded words in inverted index and eventually sends back the matched files in the rank sequence according to total relevant score.

4. Results and Discussion

The proposed scheme implements cost effective search scheme over encrypted private cloud data in this pay-as-per-use cloud paradigm. It not only returns the exactly matched files but also the files which include terms which are semantically related to the query keyword. Thus, it reduces overhead on the data user.

5. Conclusion

Hence, in this paper an attempt is made to solve the problem of efficient rank search over encrypted private cloud data. Here, we strengthen the security factor by using asymmetric encryption algorithm. Thus, the proposed scheme is secure and privacy preserving while it correctly realizes the goal of ranked keyword search.

References

[1] Xia et al.: Secure semantic expansion based search over encrypted cloud data supporting similarity ranking. Journal of Cloud Computing: Advances, Systems and Applications 2014 3:8.

[2] Prof C. R. Barde, “Secured Multiple-keyword search over encrypted Cloud data,” International Journal of Emerging Technology and Advanced Engineering, Volume 4,Issue 2,Feb 14.

[3] Xingming Sun, Yanling Zhu, Zhihua Xia and Lihong Chen, “Privacy preserving keyword based Semantic Search over Encrypted cloud data,” International journal of Security and its Applications, Volume 8, No.2 (2014).

[4] Zhangjie Fu, Xingming Sun, Senior, Nigel Linge, Lu Zhou, “Achieving Effective Cloud Search Services: Multi-keyword Ranked Search over Encrypted Cloud Data Supporting Synonym Query,” IEEE Transactions on Consumer Electronics, Vol. 60, No. 1, February 2014.

[5] Song DX, Wagner D, Perrig A, “Practical techniques for searches on encrypted data,” Proceedings of IEEE Symposium on Security and Privacy.IEEE, Berkeley, California, pp 44–55(2000)

[6] Cao N, Wang C, Li M, Ren K, Lou W, “Privacy-preserving multi-keyword ranked search over encrypted cloud data,” Proceedings of IEEE INFOCOM. IEEE, Shanghai, China, pp 829–837(2011)

[7] Yang C, Zhang W, Xu J, Xu J, Yu N “A Fast Privacy-Preserving Multi-keyword Search Scheme on Cloud Data,” International Conference on Cloud and Service Computing (CSC). IEEE, Shanghai, China, pp 104–110(2012)

[8] Stefanov E, Papamanthou C, Shi E, “Practical Dynamic Searchable Encryption with Small Leakage,” NDSS ’14, San Diego, CA, USA

[9] Liu C, Zhu L, Li L, Tan Y, “ Fuzzy keyword search on encrypted cloud storage data with small index,” IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS). IEEE, Beijing, China, pp 269–273(2014)

[10] Wang C, Ren K, Yu S, “Urs KMR Achieving usable and privacy-assured similarity search over outsourced cloud dat,” Proceedings of IEEE INFOCOM. IEEE, Orlando, Florida, USA, pp 451–459(2012)

[11] Li J, Wang Q, Wang C, Cao N, Ren K, Lou W, “Fuzzy keyword search over encrypted data in cloud computing,” Proceedings of IEEE INFOCOM. IEEE, San Diego, CA, USA, pp 1–5(2014)