ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 8, Issue 3, March 2021

Privacy Preserving Keyword Search by Index Confidentiality

[1] Sampada K S, [2] N P Kavya

[1][2] Department of CSE, RNS Institute of Technology, VTU, Bengaluru, India
Email: [1] k.s.sampada@rnsit.ac.in, [2] npkavya@gmail.com

Abstract*--- Data owners are encouraged to outlet their data to the cloud for greater flexibility and suitability. Cloud storage is a third party server which always raises question of Privacy and confidentiality. Most of the cloud servers extend to provide confidentiality by encrypting the data before outsourcing. This demands verifiability of the encrypted files that are retrieved from the cloud by the users using Query keyword search. This paper focuses on study of various search techniques over encrypted data and the threat models posed by each one of them, and then enhances on building index confidentiality for Privacy Preserving Keyword Search (PPKS). An improved privacy preserving keyword search scheme over encrypted data is being proposed to address this problem. To enable users to search over encrypted data, we use the tree structure with m-levels to build search index. The index is encrypted using secret key which is generated using random numbers. The tree is split into levels based on these secret key factors. Thus the index vectors are encrypted. Query vectors are also encrypted in the same manner and the similarity measure is considered to find the relevant document. The documents which have the highest similarity measure are considered to be relevant. The proposed model in this paper addresses the issue on *known cipher text attack.

***Keywords*--- Query keyword Search, Privacy Preserving keyword Search, Index Confidentiality, Known cipher text attack, sChosen plain text attack**

I. INTRODUCTION

Cloud computing has become inevitable and huge amount of data is being pushed on to cloud for easy access and reduced management cost. Despite the popularity and unlimited access to storage, privacy preserving is one of the essential demands that need to be addressed. Cloud Service providers (CSP) usually enforce users' data confidentiality by deploying intrusion detection systems, firewalls and virtualization. Since CSPs have full control over the infrastructure they may fail to protect the privacy of the users' data. Although Encrypting-before-outsourcing [1] can preserve privacy against CSPs, encrypted data makes data utilization a very challenging task. Query keyword search is the main objective of data utilization, is to obtain users' information of interest from huge collection of data. The cloud will become a mere storage without usable data services and can provide very limited services to the party. Hence computation over encrypted data is challenging. Both Asymmetric and Symmetric key cryptographic algorithms can be used to build encrypted data search schemes. Asymmetric encryption schemes are computationally intensive but supports more flexible search function, while Symmetric encryption schemes are more efficient but less flexible.

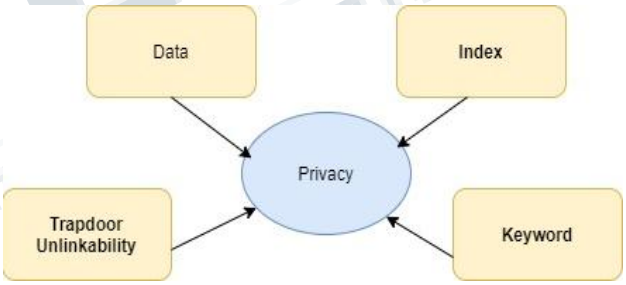

Fig.1. Requirements for Protecting Privacy

PPKS necessitate the following requirements as shown in Fig 1.

Data privacy- Data Stored in the cloud by the owner is to be protected by the unauthorized users from the CSP.

Index privacy- Index is a way of organizing the documents for faster retrieval which is also stored in the cloud and needs to be protected from the unauthorized users.

Keyword privacy: Privacy preservation is one of the trivial concern to protect the keywords of interest from the users' against CSPs. That is CSPs should not be able to infer the details of users' interest. Although the trapdoor generation can be performed using any cryptographic

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 8, Issue 3, March 2021

algorithms, CSPs can identify the keywords of users' interest by side channel attacks. This is called as Predicate privacy. Predicate encryption is a new encryption paradigm which gives a master secret key owner fine-grained control over access to encrypted data. Asymmetric encryption schemes fail to protect the predicate privacy.

Trapdoor Unlinkability: It is easy for the attacker to determine the relationship given any 2 trapdoor. Therefore it is necessary that the trapdoor should have some randomness. Hence sufficient non-determinism should be introduced into the trapdoor generation.

II. RELATED WORK

In this section, the various privacy-preserving keyword search schemes are discussed and compared based on the number of keywords, number of users along with their threat model, Privacy, Verifiability [to verify the correctness of the search result] and Efficiency. It is as shown in table 1. The mentioned techniques in the table fail to address an efficient privacy preserving multi-keyword search over encrypted data.

Table 1. Comparison of various PPKS schemes

Author	Search Technique	Methodology	Threat Model	Privacy	Verfiability	Efficiency
[2]	One keyword	One Owner and One User	Honest-but Curious	Data, index and trap Door	No	No
[3]	Ranked Multikeyword	Multi Owner and Multi User	Curious-but-Honest	Data and keyword	No	No
[4]	Ranked Multikeyword	One Owner and One User	Known Ciphertext and chosen plain text	Data, index,keyword, trapdoor	No	Yes
[5]	Multikeyword	One Owner and One User	Honest-but Curious	Keyword and Data	Yes	Yes
[6]	Similarity keyword	Multi Owners and Multi users	Honest-but-curious	Keywords and Documents	No	Yes
[7]	Smart Semantic Search	One Owner and One User	Known Ciphertext and chosen plain text	Data,Index, Keyword and Trapdoor	No	No
[8]	Ranked Fuzzy keyword Search	One Owner and One User	Honest-but-curious	Data and Index	No	Yes

III. PROBLEM FORMULATION

In this paper we propose a methodology to address privacy preserving keyword search over encrypted data by constructing index confidentiality. In this paper we also explore the design to thwart the known cipher text attack. The architecture of the PPKS over encrypted data in cloud is as shown in Fig 2. It consists of 3 entities. Data owner is the one who uploads the documents in an encrypted form along with its possible keywords which is structured as tree of m-levels.

Data user is the one who has been authenticated by the data owner to search for the encrypted data. Cloud server is the third party server where data is being stored.

The typical threat model most of the search schemes implement is to consider the cloud server to be "honest-but-curious" that is the cloud server is honest in implementing the protocol but it is curious to deduce and investigate the data by keeping an insight into the flow of messages. Since the term frequency (tf) is used to index the documents, in known cipher text attack, the attacker may extract statistical information. Given this information, the cloud server is able to launch statistical attack to deduce the keyword in a query.

In this paper we address the Known Ciphertext attack and chosen plain text attacks as a threat model from the CSP assuming that the cloud server acts as " honest-but curious" manner.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 8, Issue 3, March 2021

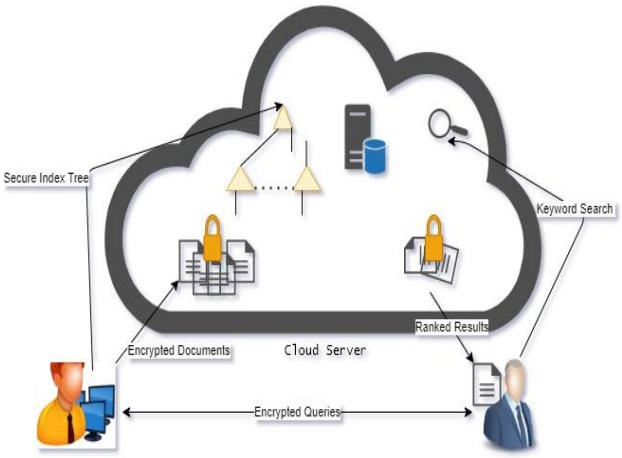

Fig 2: Architecture of PPKS over encrypted data in cloud

IV. METHODOLOGY

To enable PPKS over encrypted data over cloud the following methodologies have been considered:

i. To achieve multi-keyword ranked search, we consider representing the documents in a vector space model with cosine similarity measure.

ii. To improve the search efficiency a new tree based index structure with confidentiality is explored.

iii. In order to protect user privacy from CSP the Index confidentiality is addressed.

Vector-Space: Vector space model is one of the accepted similarity measures in the information retrieval. It holds up for both conjunctive and disjunctive search. Ranking order is usually measured as COSINE values which show the deviations of angles among the document vector and the query vector. The documents are preprocessed by removing stopwords and punctuations and are represented as vectors using tf-idf [term frequency-inverse document frequency].

Index confidentiality: The index formed from the documents is divided into sub vectors representing subset of keywords which is represented as the ith level of the index tree as shown in Fig 3. Each keyword represents an attribute domain. The attributes which has the same value forms a child nodes.

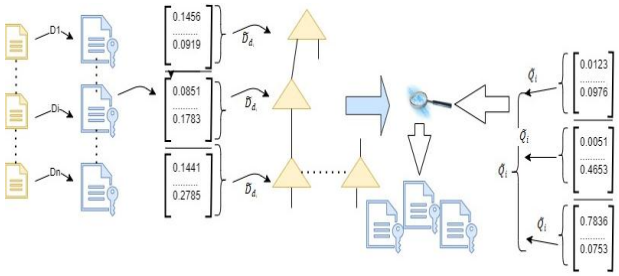

Fig 3: Architecture of index confidentiality

An index vector $D_{d,i}$ generated for each document *D* is according to subset of keywords and each dimension is a normalized term frequency weights. Each $D_{d,i}$ is split into 2 random vectors as $\{D^{x}_{d,i}, D^{xx}_{d,i}\}$, using splitting factor S_i which is a Secret key generated by the data owner by randomly generated vector S_i. M_1 and M_2 are randomly generated invertible matrices. Finally encrypted index vector $\tilde{D}_{d,i}$ is built as $\left\{M^{T}_{1.i}D^{x}_{d,i}, M^{T}_{2,i}D^{xx}_{d,i}\right\}$

Q_i a query vector is also split into 2 random vector as $\left\{Q^{x}_{i}, Q^{xx}_{i}\right\}$ with similar splitting procedure. The encrypted query vector $\tilde{Q}_i$ is yielded as $\left\{M^{-1}_{1,i}Q^{x}_{i}, M^{-1}_{2,i}Q^{xx}_{i}\right\}$.

Each level provides the similarity score of the index and the query keyword. The scores from each level are aggregated to find the final score. Cloud server considers the final similarity score to find the relevance of document to the query keyword. Similarity score of the ith level is measure as

$$\cos(\tilde{D}_{d,i}.\tilde{Q}_i) = \left\{M^{T}_{1.i}D^{x}_{d,i}, M^{T}_{2,i}D^{xx}_{d,i}\right\}.\left\{M^{-1}_{1,i}Q^{x}_{i}, M^{-1}_{2,i}Q^{xx}_{i}\right\}$$
$$= D^{x}_{d,i}Q^{x}_{i} + D^{xx}_{d,i}Q^{xx}_{d,i}$$
$$= D_{d,i}.Q_i \quad (1)$$

Where $\tilde{D}_{d,i}.\tilde{Q}_i$ represent index vector and query vector respectively in encrypted form.

Since $\tilde{D}_{d,i}.\tilde{Q}_i$ are encrypted and cannot be inferred by the Cloud server as long as S_i is confidential. The cloud server cannot deduce keywords nor the TF-IDF information from the documents or the query. Hence the known cipher text attack can be avoided in this model.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 8, Issue 3, March 2021

V. RESULTS DISCUSSION

In this section we discuss the privacy preserving keyword search by implementing index confidentiality by showing that the search over index with encryption and without encryption results to the same document set. This experiment is carried out on Google Colab. Here we have considered a sample data set consisting of 15 words which can be scalable. Following is the dataset

data type is a variable which can occupy memory allocati on and can be operable

The first step of data preprocessing is done by lemmatization and stopword removal. The figure 4 shows the data after preprocessing.

```
PREPROCESSED DOCUMENT: data type variable occupy memory allocation operable
```

Fig 4: preprocessed data

The index will be created using bag of words generated from preprocessing and their term frequencies which are normalized. The index is then split into 2 vectors based on random splitting factor. Two random matrix [Mi] are generated from split factor which are considered to be the secret key as shown in figure 5. Here the split factor is 5 which is randomly generated.

```
M1= [[0.56781468 0.24497016 0.39759231 0.14465915 0.05495402]
 [0.94085115 0.10372483 0.92038815 0.75966746 0.6594677 ]
 [0.27927849 0.79888118 0.37192899 0.30465635 0.24641766]
 [0.96231534 0.13282629 0.54996367 0.34384638 0.02190149]
 [0.67383372 0.33014708 0.5706582  0.14820139 0.14376285]]
M2= [[0.24091337 0.18633263]
 [0.98253661 0.38958511]]
```

Fig 5: Secret key matrices

These Matrices are used to create the encrypted index by multiplying with the index vector which is as shown in figure 6.

encrypted index 1: [0.06584795 0.03097211 0.05404868 0.03271213 0.02166353]

encrypted index 2: [0.02352788 0.01107534]

Fig 6: Encrypted Index

The user enters the Query keywords to search, which in turn is spit into 2 vectors with the same splitting factors. These query vectors in turn are multiplied with inverse of the random matrices respectively.

```
QUERY: data type
ENCRYPTED QUERY 1: [0.01923077 0.01923077 0.         0.         0.         ]
ENCRYPTED QUERY 2: [0. 0.]
```

Fig 7: Encrypted Query Index

The encrypted vector provides non-deterministic encryption as the splitting process includes randomization. Hence the same search keywords are encrypted to different query vectors. Hence Trapdoor unlinkability can be achieved.

The above model still suffers from chosen plain text attack as the similarity measure between the document vector and the query vector is equal to the encrypted vectors of the same as illustrated in figure 7.

```
searching over index with encryption : 0.0007396449704142011
searching over index without encryption : 0.0007396449704142013
```

Fig 7: Search Result

To protect from the keyword privacy breach, we introduce phantom terms into the query vector and increase the size of the index vector by the phantom terms. If the cloud server executes similarity evaluation it obtains the measure with phantom terms which is different from the actual measures. The phantom terms are to be selected carefully such that the corresponding distribution is not keyword specific. Hence chosen plain text attack can be avoided.

VI. CONCLUSION

With the advent of cloud computing more and more sensitive data are uploaded to the cloud server to reduce the management cost and to have ubiquitous access. The outsourcing of data introduces serious challenges as there may be an intrusion into the data that have been stored. In this paper we focused on analyzing various privacy preserving keyword search techniques performed over encrypted cloud data and compared them based on their methods, number of users and owners and Threat model posed by them. Methodology varies from single keyword, multikeyword and ranked multi-keyword. The techniques are also compared for the threat model which are mainly known cipher text and chosen plain text. Then we discussed a novel privacy preserving keyword search scheme to address the challenges mentioned with the threat model. Finally, with the proposed search model we have shown that the known cipher text attack can be avoided by implementing index confidentiality using secret key approach. But the model still suffers from chosen plain

text attack. The discussed secure search system is efficient enough to be deployed in practice. In this paper we also have not focused on access pattern which can be used for statistical attack by the CSP.

REFERENCES

[1] Boneh, Dan, Giovanni Di Crescenzo, Rafail Ostrovsky, and Giuseppe Persiano. "Public key encryption with keyword search." In International conference on the theory and applications of cryptographic techniques, Springer, Berlin, Heidelberg, 2004, pp. 506-522.

[2] Jiang, Xiuxiu, Jia Yu, Fanyu Kong, Xiangguo Cheng, and Rong Hao. "A novel privacy preserving keyword search scheme over encrypted cloud data." In 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), IEEE, 2015, pp. 836-839

[3] Zhang, Wei, Yaping Lin, Sheng Xiao, Jie Wu, and Siwang Zhou. "Privacy preserving ranked multi-keyword search for multiple data owners in cloud computing." IEEE Transactions on Computers Vol. 65, No. 5, 2015, pp. 1566-1577.

[4] Gurjar, Sonu Pratap Singh, and Syam Kumar Pasupuleti. "A privacy-preserving multi-keyword ranked search scheme over encrypted cloud data using MIR-tree." In 2016 International Conference on Computing, Analytics and Security Trends (CAST), IEEE, 2016, pp. 533-538.

[5] Wan, Zhiguo, and Robert H. Deng. "VPSearch: achieving verifiability for privacy-preserving multi-keyword search over encrypted cloud data." IEEE transactions on dependable and secure computing 15, no. 6 (2016): 1083-1095.

[6] Li, Jinguo, Mi Wen, Chunhua Gu, and Hongwei Li. "PSS: Achieving high-efficiency and privacy-preserving similarity search in multiple clouds." In IEEE International Conference on Communications (ICC), IEEE, 2016. pp. 1-6.

[7] Fu, Zhangjie, Fengxiao Huang, Kui Ren, Jian Weng, and Cong 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, 2017, pp. 1874-1884.

[8] Xu, Qunqun, Hong Shen, Yingpeng Sang, and Hui Tian. "Privacy-preserving ranked fuzzy keyword search over encrypted cloud data." In International Conference on Parallel and Distributed Computing, Applications and Technologies, IEEE, 2013, pp. 239-245