ISSN:0975-9646

An Efficient and Privacy Preserving in Multi-Keyword Ranked Search over Encrypted Cloud Data

Jaikishan A Tindwani[#1], Prof. Aruna K Gupta[*2]

[#]*Computer Science Department, JSCOE, Pune University*
Pune, Maharashtra, India

[*] *Information Technology Department, JSCOE, Pune University*
Pune, Maharashtra, India

***Abstract*— Innovation in cloud computing have revamped the view of modern information technology which is motivates the data owners to outsource the data to public cloud server for fast access to data its management at minimum cost. Earlier it was not possible to upload the encrypted data over the cloud. Now-a-days with the increasing number of users the data is also increasing, so there is need to provide security to data over the cloud. To provide security the document should be first encrypted before outsourcing it and it can be retrieved effectively. Some more additional features can be provided along with search such as dynamic update operations. For the index creation and query generation the vector space model and the TF IDF model are combined. A tree-based index structure is constructed and a GDFS algorithm is proposed to produce efficient multi-keyword ranked search. kNN algorithm is used for encrypting the index and query vectors, to acquire accuracy of relevance score between encrypted index and query vectors. To avoid statistical attacks, phantom terms are included to index vector for building the search results.**

***Keywords* - Encrypted Data Search, Cloud Service Providers, Cloud Storage, Ranked Search.**

I. INTRODUCTION

Cloud computing is nowadays a widely used technology for providing services over the network. But as the usage of cloud storage is increasing, the security risk, integration of data, and its confidentiality are also implicitly increasing. Therefore the cloud service providers should manage security and confidentiality, as they are playing a vital role in data sharing functionality. The special care should be taken care for the data security, as it is storing at cloud storage which is managed by third party. As the usage of internet is increasing, the users prefers to store/upload data on cloud so that they can access the data from anywhere in the world. But keeping the privacy in mind, traditional data storage techniques for authentication are not that much reliable. For the protection of the data over cloud, the data needs to be encrypted before uploading them to cloud to avoid the escalation which may open up with the confidentiality of data. Cloud storage is the very important and widely used cloud computing model where data is stored on remote servers and managed and accessed over internet. It is managed and operated by the CSP on a server which support data storage and is built on virtual machines. It works through the data centre virtualization which provides applications and data users a virtual architectural environment that is scalable according to its requirements. All the search schemes which support the multi-keyword functionality retrieve search output.

II. LITERATURE SURVEY

Paper Name: A Secure and Dynamic Multi-keyword Ranked Search Scheme over Encrypted Cloud Data [1]
Author: Zhihua Xia, Xinhui Wang, Xingming Sun, Qian Wang
Description: In [1], author describes a secure multi-keyword ranked search scheme over encrypted cloud data, which supports dynamic update operations like deletion and insertion of documents. The vector space model and the widely-used TFIDF model are combined in the index construction and query generation. A special tree-based index structure and introduces a Greedy Depth-first Search algorithm to provide efficient multi-keyword ranked search. The secure kNN algorithm is utilized to encrypt the index and query vectors, and to ensure accurate relevance score calculation between encrypted index and query vectors.

Paper Name: Enabling Secure and Efficient Ranked Keyword Search over Outsourced Cloud Data [2]
Author: Cong Wang
Description: In [2], author proposed search which solves processing overhead, data and keyword privacy, minimum communication and computation overhead. The owner build index along with the keyword frequency-based relevance scores for files. User request 'w' to CS with optional 'k' as Tw using the private key. The CS searches the index with scores and sends encrypted file based on ranked sequence.
Advantage: The CS searches the index with scores and sends encrypted file based on ranked sequence.
Disadvantage: It does not perform multiple keyword searches and little overhead in index building.

Paper Name: Privacy-Preserving Multi Keyword Ranked Search over Encrypted Cloud Data [5]
Author: Ning Cao
Description: In [5], proposed this search for known cipher text model and background model over encrypted data providing low computation and communication overhead. An Efficient and privacy preserving in Multi-Keyword Ranked Search over

encrypted cloud data the coordinate matching is chosen for multi-keyword search. They used inner product similarity to quantitatively evaluate similarity for ranking files. The drawback is that MRSE have small standard deviation ← which weakens keyword privacy.

Paper Name: Efficient and Secure Multi-Keyword Search on Encrypted Cloud Data [6]
Author: Li, S. Yu
Description: In [6], defined and solved the problem of effective but safe and sound rank keyword search over Encrypted cloud data. Ranked search greatly enhances system usability by returning the matching files in a ranked order regarding to certain relevance criteria (e.g., keyword frequency) thus making one step closer towards sensible consumption of privacy preserving data hosting services in Cloud Computing. These papers has defined and solved the challenging problem of privacy-preserving multi-keyword ranked search over encrypted cloud data (MRSE), and establish a set of strict privacy requirements for such a protected cloud data utilization system to become a reality. The proposed ranking method proves to be efficient to go back extremely relevant documents corresponding to submitted search terms. The idea of proposed ranking method is used in our future system in order to enhance the security of information on Cloud Service Provider.

Paper Name: Single Keyword Search over Encrypted data on cloud [3]
Author: Madane S.A.
Description: In [3], Obtainable searchable encryption scheme consent to a user to firmly look for over encrypted data through keywords without first decrypting it, these techniques support only conventional Boolean keyword search, without capturing any relevance of the files in the search result. When directly applied in large joint data outsourcing cloud environment, they go through next shortcoming.
Advantage: Support only conventional Boolean keyword search without decrypting it.
Disadvantage: Single-keyword search without ranking, Boolean- keyword search without ranking and Do not get relevant data.

III. Problem Definition

To provide effective multi-keyword ranked search on encrypted data over cloud supporting dynamic update operations and to minimize the search time by building the special tree-based index structure using ”Greedy Depth-first Search” algorithm.

IV. Mathematical Modelling And Algorithm

Defined Notations:
W - The dictionary, namely, the set of keywords, denoted as W = w1; w2… wm
m - The total number of keywords in W
Wq - The subset of W, representing the keywords in the query
F - The plaintext document collection, denoted as a collection of n documents F = f1; f2… fn. Each document f in the collection can be considered as a sequence of keywords.
n - The total number of documents in F.
C - The encrypted document collection stored in the cloud server, denoted as C = c1; c2… cn.
T - The unencrypted form of index tree for the whole document collection F.
I - The searchable encrypted tree index generated from T.
Q - The query vector for keyword set Wq.
TD - The encrypted form of Q, which is named as trapdoor for the search request.
Du - The index vector stored in tree node u whose dimension equals to the cardinality of the dictionary W. Note that the node u can be either a leaf node or an internal node of the tree.
Iu - The encrypted form of Du.

Set Theory:-

Let S be the system to perform Multi Keyword Ranked Search over the Encrypted Cloud Data.
S= {I, O, F, Fail, Success}
Where, Inputs I = {I1, I2, I3, I4, I5}
I1= User Login,
I2= Extract Keywords,
I3= Encryption of data using public key,
I4= File and Keywords stored on cloud,
I5= Search over the cloud data using the keywords,
I6= Decryption of data
Where, O is an Output
O = {O1, O2, O3, O4}
O1= Account Created,
O2= Keyword is Extracted in Encrypted form,
O3= Keywords are used for Encryption,
O4= Updation in the index,
O5= Encrypted data,
O6= Extraction of the original data.
Where, Functions set F
F1 {Setup},
F2 {KeyGen},
F3 {Encrypt},
F4 {Upload},
F5 {Keyword Search},
F6 {Decrypt}

Failure Conditions: In this case, the search is not successful i.e., the user is not able to find the required document because there exists no keyword matching to the required query as required by the user.

Success Conditions: In this case, the search is successful i.e., the user is able to find the required document as the keyword exists in the index; the document is retrieved by the user.

Graph Representation:-

Let G be a closed graph that represents our system; Such that G = {E, V}

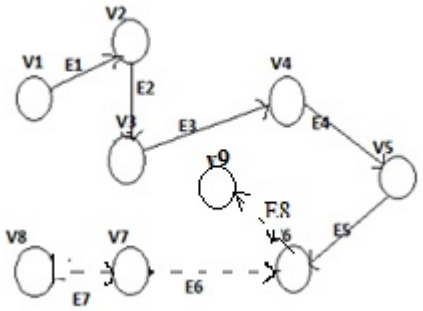

Fig 1: Graph Representation

Where,
E represents the set of edges; E = {e1, e2, e3 ....., e10} and V is a set of vertices; V = {v1, v2... v5}.

In graphical representation of system, vertices in set V represent the modules which are connected through the directed edges in set E which represents the input/output of modules.

Let fe be a rule of E into V such that for a given edge; fe will return vertices. fe (E) V.
Thus, for this system,
fe (e1) = v2.....v2 is called using e1 for keyword generation and file encryption.
fe (e2) = v3..... data is passed to v3 using e2 to upload a file.
fe (e3) = v4..... e3 is used to pass data to v4 for encryption of the keywords.
fe (e4) = v5..... e4 is used to upload the encrypted keywords and generating the index at v5.
fe (e5) = v6.....the edge e5 is passed to v6 to provide encrypted index for searching.
fe (e6) = v7.....the edge e6 is passed to v6 providing the encrypted keywords for searching.
fe (e7) = v8..... the edge e7 is passed to v7 providing the unencrypted keywords provided by the user

ALGORITHMS
Algorithm: Buildtree-Index

Input: Document collection F = {f1; f2... fn} with the identifiers ID = {ID|ID = 1; 2... n}.

Output: Index tree T
1: for each document fID in F do
2: Construct leaf node v for fID, with v.ID = GenID (), v.Pl = v.Pr = null, v.ID = ID, and D[i] = TFfID; wi for i = 1... m;
3: Add v to CurrentNodeSet;
4: end for
5: while number of nodes in CurrentNodeSet is greater than 1 do
6: if number of nodes in CurrentNodeSet is even, i.e. 2u then
7: for each pair of nodes v' and v" in CurrentNodeSet do
8: Generate parent node v for v' and v", with v.GID = GenID(), v.Pl = v', v.Pr = v", v.ID = 0 and D[i] = maxfv'.D[i]; v".D[i]g for each i = 1;.....;m;
9: Add v to TempNodeSet;
10: end for
11: else
12: for each pair of nodes v' and v" of former (2u - 2) nodes in CurrentNodeSet do
13: Create parent node v for v' and v";
14: Add v to TempNodeSet;
15: end for
16: Generate parent node v1 for the (2u - 1)-th and 2u-tu node, and then generate parent node v for v1 and the (2u + 1)-tu node;
17: Add v to TempNodeSet;
18: end if
19: Replace CurrentNodeSet with TempNode-Set and then clear TempNodeSet;
20: end while
21: return only node left in CurrentNodeSet, namely, root of index tree RT ;

V. IMPLEMENTATION STRATEGY

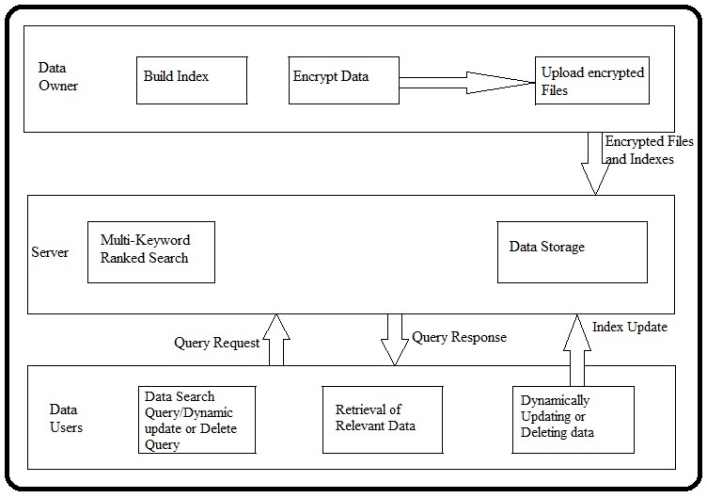

Fig 2: System Architecture

1. Here, the data owner will encrypt the file using the external application. Along with the encrypted file the user (Data Owner) will get the keywords which he/she can use for the search purpose.

2. After that, the encrypted file gets uploaded to the server. Encrypted keywords get uploaded to the server also.

3. Index will get generated using the build tree Index algorithm.

4. End user is able to search the data but cannot see the encrypted data without the permission of data owner.

5. Request will be sent to the owner that the following user wants to access the content of the file.

6. Data owner will then send the key to the user privately through the mail.

7. End user has the ability to change the content of the file and save with different version.

The system consists of 4 main implementation steps:-

SK Setup (): Here system set the secret vector S as an m-bit vector, and set M1 and M2 are (m + m') (m + m') invertible matrices, where m is the number of phantom terms.

I GenIndex (F; SK): Before encrypting index vector Du, extend the vector Du to be (m+m') - dimensional vector. Each extended element Du [m+ j], j = 1... m', is set as a random number E (j).

TD GenTrapdoor (Wq; SK): Query vector Q is extended to be (m + m') - dimensional vector. Among the extended elements, a number of m elements are randomly chosen to set as 1, and the rest are set as 0.

RelevanceScore SRScore(Iu; TD): After the execution of relevance evaluation by cloud server, the final relevance score for index vector Iu equals to Du.Q + ΣE(v), where v ε {j— Q[m + j]} = 1.

VI. Results

The implementation of the proposed scheme is done using Asp.Net and C # language in Windows 7 operation system and tests its efficiency. The tests include Search precision on varied privacy level. The experimental results are produced with an Intel Core(TM) i5 Processor.

The search precision of this system is affected by the phantom keywords in proposed method. Here, the precision is defined as that in [5]: Pk = k'/k, where k' is the number of real k documents as top ranked in the retrieved k documents. If a minor standard deviation is set for the random variable − ⍰ v, this technique is supposed to obtain larger precision, and vice versa. The results are shown in Fig.3 (a). As said earlier here phantom terms are added to the index vector to change the relevance score calculation, so that the cloud server cannot detect keywords by checking the TF distributions of special keywords. Here, we quantify the obscureness of the relevance score by "rank privacy", which is defined as:

$$P'_k = \Sigma \left| r_i - r'_i \right| / k_2;$$

Fig 3 (a) Precision of searches with different standard deviation

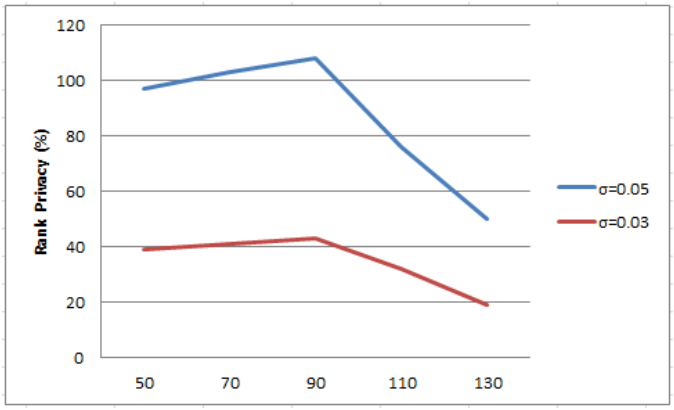

Fig 3 (b) Rank Privacy of searches with different standard deviation

Where ri is the rank number of document in the retrieved k documents as top documents, and r′ i is its real rank number in the whole ranked results. The larger rank privacy denotes the higher security of the scheme, which is shown in Fig. 3(b). Here, data users can accomplish different requirements on search precision and privacy by adjusting the standard deviation σ.

Here the comparison is with a recent work, which achieves good search efficiency. The previous scheme retrieves the search results through exact calculation of document vector and query vector. Thus, top-k search precision of the previous scheme is 91%. But as a similarity-based multi-keyword ranked search scheme, the previous scheme suffers from precision. The average precision of this method is 86%.

VII. Conclusions

The proposed method for the multi keyword ranked search on encrypted data allows not only search but also the dynamic updating information of those files. Here, the data owner is also responsible for the updating information of the information needed for the index generation. So the extra steps are required by the data owner every time he uploads the file to the cloud. But there are many challenges in a secure multi user system i.e. the data owner consider that all the users are trustworthy ad not dishonest. But this is impractical; a dishonest user may lead to many secure problems like sharing the secure key with the unauthorized person or sharing the decrypted data file with another organization. In the future works, we will try to accomplish the challenges for the secure system to improve its privacy...

Acknowledgment

I take this opportunity to express my profound gratitude and deep regards to my guide Prof. A.K. Gupta for her exemplary guidance, monitoring and constant encouragement which helped me in completing this task through various stages. The Blessings, help and guidance given by her time to time shall carry me a long way in the journey of life which I am about to embark.

REFERENCES

[1] Zhihua Xia et al,"Secure and Dynamic Multi-keyword Ranked Search Scheme over Encrypted Cloud Data", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL., NO. 1.

[2] Cong Wang et al.,"Enabling Secure and Efficient Ranked Keyword Search over Outsourced Cloud Data", IEEE Transactions on parallel and distributed systems, vol. 23, no. 8, August 2012.

[3] Madane S.A, "Comparison of Privacy Preserving Single- Keyword Search and Multi-Keyword Ranked Search Techniques over Encrypted Cloud Data", 2014 International Journal of Computer Applications (0975 - 8887) Volume 126 - No.14, September 2015.

[4] Wenhai Sun et al., "Privacy-Preserving Multikeyword Text Search in the Cloud Supporting Similarity-based Ranking", the 8th ACM Symposium on Information, Computer and Communications Security, Hangzhou, China, May 2013.

[5] Ning Cao et al.,"Privacy-Preserving MultiKeyword Ranked Search over Encrypted Cloud Data", IEEE Transactions on parallel and distributed systems, vol. 25, no. 1, jan 2014.

[6] Ming Li et al,"Authorized Private Keyword Search over Encrypted Data in Cloud Computing", IEEE proc. international conference on distributed computing systems, June 2011, pages 383-392.

[7] A. Singhal "Modern Information Retrieval: A Brief Overview", IEEE Data Eng. Bull., vol. 24, no. 4, pp. 35-43, Mar. 2001.

[8] D. Song, D. Wagner, and A. Perrig"Practical Techniques for Searches on Encrypted Data", Proc. IEEE Symp. Security and Privacy, 2000.

[9] Shih-Ting Hsu et al.,"A Study of Public Key Encryption with Keyword Search", International Journal of Network Security, Vol.15, No.2, PP.71-79, Mar. 2013.

[10] Kui Ren et al.,"Towards Secure And Effective Data utilization in Public Cloud" IEEE Transactions on Network, volume 26, Issue 6, November / December 2012.

[11] Wenhai Sun et al.,"Privacy-Preserving Multikeyword Text Search in the Cloud Supporting Similarity-based Ranking" the 8th ACM Symposium on Information, Computer and Communications Security, Hangzhou, China, May 2013.

[12] K. Wong, D. W.-l. Cheung, B. Kao, and N. Mamoulis "Secure knn computation on encrypted databases" Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. ACM, 2009, pp. 139152.