Secure Keyword Based Search in Cloud Computing: A Review

Manjeet Gupta[1], Sonia Sindhu[2]
[1]Assistant Professor, Department of Computer Science & Engineering,Seth Jai Prakash Mukund Lal Institute of Engineering &Tech, Radaur, India .
[2]M.tech pursuing, Department of Computer Science & Engineering, Seth Jai Prakash Lal Institute of Engineering &Tech, Radaur, India .

ABSTRACT- An effective search that give the better suggestion to the user so that the user can get better choices for the services is also a challenge for cloud computing. Several methods have been given to solve the problem of effective and secure ranked keyword search of cloud data. Ranked search greatly enhances system usability by returning the matching files in a ranked order regarding to certain relevance criteria , like keyword frequency. Thus making one step closer towards practical deployment of privacy – preserving data hosting services in cloud computing. In this paper we present a study on keyword search that perform the parametric matching on different cloud services and perform a rank based selection on cloud services . Here we also discuss the security provided by all the techniques.

1.INTRODUCTION

Cloud computing has become hot topic in the IT industries. Great efforts have been made to establish cloud computing platform for small enterprise user. A cloud services is the new trend of computing where readily available computing resources are exposed as a service .These computing resources are generally offered as pay-as-you -go plans and hence have become attractive to cost conscious customers. Apart from the cost ,cloud services also supports the growing concerns of carbon emissions and environmental impact because the cloud advocates better management of resources. We see a growing trend of off loading the previously in-house service systems to the cloud. Such a move allows businesses to focus on their core competencies and not burden themselves with back office operations.

Effective information management and retrieval within an enterprise or over internet to a large extent depends heavily on both the organizing , searching ,browsing and navigating facilities built in to information management systems and their intuitiveness and usability. Information search and document retrieval from a remote database requires submitting the search terms to the database holders. However search terms may contain sensitive information that must be kept secret from the database holder. Moreover the privacy concern is required to apply to the documents retrieved by the user in the later stage because they may contain sensitive information about search terms. Therefore in this paper we will study various secure and fast information retrieval methods by using user query keywords.

3. RELATED WORKS

Keyword search have been implemented in various databases to get the better result. One of the major requirement over the web is about the selection of best service and service provider over the web[1]. When we talk about cloud service the work is more specific and the parametric. The tag cloud is also used to search the relevant information on the basis of tag assignment to different kind of keywords and on the basis of these tags a query refinement is been performed. Finally a flexible search over the database is performed to derive the final outcome. The result analysis is based on the basis of effectiveness and efficiency of the cloud services[2]. Every keyword search follow some common step which can be explained below:

Keyword Extraction: A word used by search engine in its search for relevant web pages. In this architecture at first the query is performed by the user and on this query the query analysis is performed. The analysis includes the keyword extraction by removal of stop words. Once the keyword extracted the next work is about to perform the keyword summarization based on frequency of keywords. Once we get the summarized keywords it will be used as the content based analysis.

Indexing: Another program called indexer ,reads the documents and creates the index based on the words contained in each document. In information, indexing structure is used to store a list of mappings from

keywords to the corresponding set of files that contain this keyword allowing full text search[3].

Ranking formula: For ranked search purposes ,the job of determining which files are most relevant is determined by assigning a numerical score ,which can be determined to each file based on some ranking formula. Ranking function is used to calculate relevance scores of matching files to a given search request. The most widely used formula for evaluating relevance score in the information retrieval is TF*IDF , where TF (term frequency) is simply the number of times a given term (keyword) present within sa file to represent the importance of term within a file.IDF (inverse document frequency) is given by dividing the number of files in the whole collection by number of files containing the keyword to measure the overall importance of the keyword in the whole collection. Now we will explain different methods which uses these above discussed steps in information retrieval. We have to search the files as well as maintain the privacy of the documents retrieved and stored at the server. Several protocols have also used to maintain the security in ranked search in cloud computing. One of them is private information retrieval(PIR) ,provides useful cryptographic tools to hide the queried terms and the the data retrieved from the database while returning most relevant documents to the user. With the growth of music collection ,music information retrieval (MIR) has been given in recent years. There are several ways to retrieve pieces of desired music. For example query by meta-information and query by tag. In content based MIR system ,user input a query of multiple tags with multiple level of preference by colorizing desired tags in a web based tag cloud interface to search music[4].Keyword search of PubCloud is also used in PubMed (database of biomedical literature). PubMed which is part of National Center for Biotechnology Information (NCBI), is a centralized database that indexes millions of biomedical publications. Responses to queries are presented by ranked list that are similar to responses of most web search engines[6].

1) Public Key Encryption With Keyword Searching: It is asymmetric searchable encryption scheme ,where encryption is done using a public key system. PEKS is said to be first predicate encryption scheme.The underlying scheme is that , when a user requests a particular keyword ,the server should not know anything about the keyword or document. The four algorithms used in this technique are following:

a)Key Generation :Takes a security parameter ,s,and generates a public or private key pair A_{pub} , A_{priv}.

b)PEKS(A_{pub},W): For a public key A_{pub} and a word W produces a searchable encryption of W.

c)Trapdoor($A_{priv,}$W): With A's private key and a word W , a trapdoor Tw is produced.

d)Test(A_{pub},S,T_w): With the public key of A, searchable encryption S=PEKS($A_{pub,}$W')

and the trapdoor Tw=Trapdoor(A_{priv},W) outputs yes if W=W else no.

After generating the key for the user and the server ,the PEKS algorithm is executed. After receiving the keyword the trapdoor is calculated. After trapdoor generation the test is conducted to see whether the given keyword matches the one requested.It is proven semantically secure against the adaptive chosen keyword attack. The disadvantage of this method is that ,when multiple keywords are used , the search is not possible. [1]

2)Efficient and Secure Ranked Multi-Keyword Search on Encrypted Cloud Data: A related protocol , Private Information Protocol(PIR),provides useful cryptographic tools to hide the queried search terms and the data retrieved from the database while returning most relevant documents to the user. This practical privacy-preserving ranked keyword search scheme allows multi-keyword queries with ranking capability. Here three roles have considered:

Data owner, who is the actual owner of the database. The data owner collects and generates the information in the database .Users are the members in a group who are entitled to access the information of the database.

Server ,is a professional entity to offer information services to authorized users. Following steps are carried out to access the information:

1)The data owner creates a search index for each document . The search index file is created using a secret key based trapdoor generation function where secret keys are only known to data owner.

2) Data owner upload these search index files together with encrypted documents. This scheme use symmetric-key encryption as encryption the encryption method since it can handle large document.

3)When a user wants to perform a keyword search , he first connects to data owner and learns the trapdoor for the keywords he wants to search for , without revealing the keywords he wants to search.

4)After getting the trapdoor information , user generates the query and submit it to the server.

5)Server sends metadata to the user and then user retrieves the encrypted documents he chooses after analyzing the metadata that basically conveys a

relevancy level of the each matched document where the number of documents returned is specified by the user.

6)Finally the user interacts with owner in order to decrypt the documents and get the corresponding plain text.

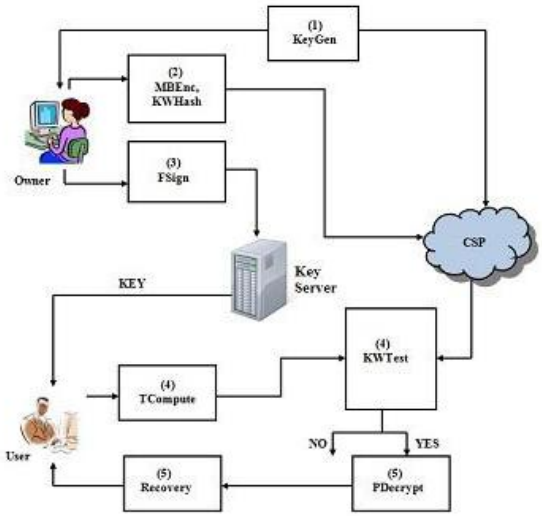
This method maintain the data privacy, index privacy , trapdoor privacy and non-impersonation.[5]

3)Fuzzy Keyword Search over Encrypted Data in Cloud Computing:

The traditional searchable encryption allows a user to securely search over encrypted data through keywords and selectively retrieve files of interest ,but these techniques allows only exact keyword search. That is there is no tolerance of minor typos and format inconsistancies which ,on the other hand ,are typical user searching behavior and happen very frequently. This drawback makes existing techniques unsuitable in cloud computing as it greatly affects system usability ,rendering user searching experiences very frustrating and system efficacy very low. Here effective fuzzy keyword search over encrypted cloud data solve this problem while maintaining keyword privacy. Fuzzy keyword search greatly enhances system usability by returning the matching files when use's searching inputs exactly match the predefined keywords or return the closet possible matching files based on keyword similarity semantics ,when exact match does not exists. The edit distance to quantify keywords similarity and developed an advanced technique on constructing fuzzy keyword sets, which greatly reduces the storage and representation overheads . There are several techniques to quantitatively measure the string similarity. The edit distance ed(w1,w2) between two words w1 and w2 is the number of operation required to transform one of them in to another .The fuzzy keyword set can be constructed by using wildcard based fuzzy set construction. In this approach ,all the variants of the keywords have to be listed even if the operation is performed at the position .The fuzzy keyword search is secure and privacy preserving . [6]

4)Efficient and Privacy Preserving Multi User Keyword Search *for Cloud Storage Services:*

The efficient and privacy preserving multi user keyword search for cloud storage sevices is also available. In this scheme ,the service provider participate in the partial decipherment of the cipher text, thus reducing the computational overhead of the user. The same keywords are encrypted to different cipher text thus reducing redundancy and avoiding the chance of statistical attack on keyword cipher text. The scheme also support the multiple users . The user who searches for document may be different from the users who encrypt and store it in cloud. User authorization is also provided by this technique. User authorization is achieved by using the digital signature of the file name. This scheme uses two server one is key server and other is cloud service provider.



(working process of above scheme)

This scheme consists of eight randomized polynomial time algorithm which are following:

i)KeyGen: It produces public private key pair for a user and public private key pair for the CSP.

ii)MBEnc: It is a public key encryption algorithm that takes two public keys of user and CSP and a message m as inputs and produces m's cipher text.

iii)KWHash:It is a one way hash function that takes a keyword and time stamp as input and produces keyword's cipher text.

iv)FSign: It is a public key signature algorithm which takes the user secret key and a file name as input and produces the signature of the file name.

v)TCompute: It takes a keyword as inputs and produces keyword's trapdoor.

vi)KWTest: It takes time stamp ,and an encrypted keyword and a trapdoor as inputs and outputs 1 or 0 .

vii)PDecrypt: It takes private key of server and public key of user and a cipher text of document as inputs and outputs an intermediate result.

viii)Recovery: It takes a private key of user and an intermediate result as inputs , and outputs the plain text.[7]

5)Towards Authenticated and Complete Query Results from Cloud Storages: The cloud storage providers (CSP) might not be fully trusted and susceptible to be compromised. The CSP might deliberatley search only part of user data for their purpose ,or they might just be incompetent to carry

out complex search request ,which yield incorrect query replies. Therefore an authentication mechanism of query result is required to enable cloud users not only to protect the security of the data in the cloud , but also to verify the correctness of the query results from the CSPs. An efficient scheme for CSPs to provide the proof of query results and for cloud users to be assured by verifying the proof. Message Authentication Code (MAC) is a mechanism used to authenticate a message. Hash based message authentication code (HMAC) combine both the cryptographic hash function(such as MD5 and SHA) and a secret key. HMAC provide both authenticity and integrity of a message. There are three types of tampered query results possible:

1) Missing data from query results.

2) Unintended data within query result.

3) Altered data within query result**.**

To our best knowledge ,three primitive approaches have discussed aiming at providing the security of query result.

In the first scheme , the index is stored locally by the CSC (cloud storage client). Therefore it causes storage computation overhead to the CSC and single point of failure if index table is missing or broken.

The second scheme uses PEKS by moving the encrypted index table to two index clouds , and the encrypted data to one data cloud .The CSC queries each of two index clouds to obtain and compare the sets of targeted identifiers. This scheme assumes that two index CSPs are not colluding but both of them can be malicious if they can benefit from doing so.

The third scheme takes the advantages of PEKS scheme to construct searchable indexes. It also uses HMAC to ensure authentication and integrity of query result. A counter is introduced for each keyword to insure the completeness of query result.

Properties	Primitive i	Primitive ii	Primitive iii
Local storage requirement	intensive	low	low
Computation requirement	intensive	low	low
Communication requirement	low	Intensive (with index cloud)	low
Query result authenticity	Provided (after decryption)	provided (after decryption	provided
Query result completeness	provided	provided (with assumption)	provided

Comparision of three primitives[8]

4. CONCLUSION

The study through all these papers has shown that different keyword searching schemes offer a way to overcome one technique's disadvantages. The keyword searching techniques improve the security of the user query and information from server but cause an increase in the cost and response time. The cost associated with communication and computation and response time of user query have to be decreased by applying some methods.

5. REFERENCES

[1]Tritty Mamachan, Roshni .M. Thanka,"Servey on Keyword Searching in Cloud Storages .International Journal of Emerging Technology and Advanced Engineering(2012).
[2] Dimitrios Skoutas (2011)," Tag Clouds Revisited", CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK. Pp 221-230.
[3] Cong Wang, Ning Cao, Jin Li(2011)," Secure Ranked Keyword Search over Encrypted Cloud Data".
[4]Ju-Chiang Wang (2011)," Colorizing Tags in Tag Cloud: A Novel Query-by-Tag Music Search System", MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.ACM p 293-302.
[5] Cengiz Orencik (2012) ," Efficient and Secure Ranked Multi-Keyword Search on Encrypted Cloud Data", PAIS 2012, March 30, 2012, Berlin, Germany. ACM, p 186-195.
[6]JinLi,Qian Wang(2010),"Fuzzy Keyword Search over Encrypted Data in Cloud Computing",Mini-Conference at IEEE INFOCOM2010.
[7]Remya Rajan(2012),"Efficient and Privacy Preserving Multi User Keyword Search for Cloud Storage Services" ,International Journal of Advanced Technology & Engineering Research (IJATER), Volume2,Issue4,July 2012.
[8]FuKuoTseng ,S RongJaveChen(2012) "Toward Authenticated and Complete Query Results from Cloud Storages" IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communication.