Privacy Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data

Muzammil Ahmed[1], Asrarullah Khan[2]

[1]M.E Dept. of CSE, Matoshri Pratishthan Group of Institutions, Khupsarwadi, Nanded, Maharashtra
[2]Assistant Professor Dept. of CSE, Matoshri Pratishthan Group of Institutions, Khupsarwadi, Nanded, Maharashtra

Abstract – *Significant hike in amount of data is observed in past couple of years. Now, it's the era of Big Data, where Volume, Velocity & Variety are major components of data sets.*

This expansion in total data and its overall usage has inspired more to data owners to migrate their data sets to cloud. This migration is sort of outsourcing. Hence, data owners outsource their data management systems from local sites to commercial public cloud. Public cloud offers economic solutions to many problems observed and provides awesome flexibility to data owners. But as the responsibility of data owners has reduced, their risk for data security has increased proportionately. Data owners can enjoy full advantage of cloud computing only if they are able to address very real secrecy and security concerns that come with storing sensitive personal information. For real secrecy, user data and their identity should remain hidden from Cloud service provider (CSP) and to protect secrecy of data, it has be encrypted before outsourcing. This encrypted data is outsourced and blocks CSP from understanding the data. As the number of data users are significantly higher, feature for an encrypted cloud data search is of great importance. Also, by considering the large number of data users and documents in the cloud, it is also important for this search service to be capable to permit multi-keyword search query and should provide results similarity ranking to meet the effective need of data retrieval search and not often distinguish the search outcomes. In this system, we define and solve the challenging problem of secrecy-preserving multi-keyword ranked search over encrypted cloud data (MRSE), and establish a set of strict secrecy requirements for such a secure cloud data utilization system to be implemented in real.

We first propose a basic idea for the Multi-keyword Ranked Search over Encrypted cloud data (MRSE) [1]. It is based on two components, first Secure Inner Product Computation and efficient similarity measure of coordinate matching. It matches as many results as possible and captures relevance of data documents in response to the search query. Later we have explained two considerably improved MRSE schemes that accomplishes various stringent secrecy requirements [2] in two unlike threat models. Unidentified ID is assigned to the user which offers better security to the data uploaded on cloud server is done.

Key Words: Cloud computing, rated search, encryption, keyword search, searchable encryption, index

1. INTRODUCTION

Cloud computing is Software as a Service (SaaS), where cloud customers can store their data onto cloud. It offers its clients to enjoy services along with the on-demand high quality applications from pool of computing resources which are shared among multiple such clients. In layman terms, cloud computing is nothing but saving and accessing the user's data on some machine connected through internet. Earlier this data used to be stored in local machine.

The importance of cloud computing is growing and receiving rising consideration in the scientific and industrial communities [3]. It is an economic, flexible, and recognized delivery platform providing business or consumer IT services using internet. Although there are many advantages of cloud computing, risks are also attached with the same as the vital services are frequently outsourced to a third party, which causes data security and privacy, support data and service availability.

Both individual clients and different companies are getting attracted with cloud computing's great flexibility and its pocket friendly alternatives. All are shifting their data from local data sets to cloud. This uploaded data needs to be protected from any unauthorized access as it may contain personal information like Bank documents, Tax documents, ID proofs, Email threads/ backups, Photos etc. Hence, this needs to be in encrypted form. This encryption is done by the data owners before uploading it to any public cloud; this, however, finished the traditional data utilization service which was built on plaintext keyword search. The insignificant solution of downloading all the data from cloud and decrypting locally is obviously impractical, due to high bandwidth utilization. Moreover, apart from rejecting the local storage management, storing data into the outsourced storage doesn't serve any purpose unless they can be easily searched and utilized.

1.1 Literature Review

Rated search system allows search user to find most matching documents instead of resulting homogeneous results in response to search query. This ranked system shows users the most relevant search results matching his search requirement and eradicates unnecessary bandwidth utilization by sending only the most pertinent data. For secrecy, this rating system should not also leak any information. Search accuracy is also improved if user is allowed to search with multiple keywords in a single search query. Using multiple keywords restricts the number of results and returns only the matched documents. Thus multiple keywords in a search request narrow down the search results further.

Another competent similarity measure is coordinate matching [9]. It matches as many as possible & useful in multiple keyword semantics to get most relevant search results. Searchable encryption [10] [11] [12] is very helpful technique as it considers encrypted data as documents and a user can search securely through a single search and gets results as per his query.

Although Boolean keyword search [7] has been proposed in recent years, it is not adequate enough to provide satisfactory result ranking functionality.

In this paper we have been proposed to search multiple keywords rated search over encrypted cloud data (MRSE). It also preserves very strict machine level secrecy in the cloud computing model. Amongst various multiple keyword semantics, proficient similarity measure of "coordinate matching" is used as it matches all the possible significance search results. The search query is also described as vector which has a bit associated with the keyword describing whether corresponding document is attached.

2. Problem Statement

The commercial savings of outsourced storage and unlimited flexibility are encouraging both organizations and individuals to outsource their local complex data management system. To protect data privacy and combat unwanted accesses in the cloud and beyond, sensitive data, may have to be encrypted by data owners before outsourcing.

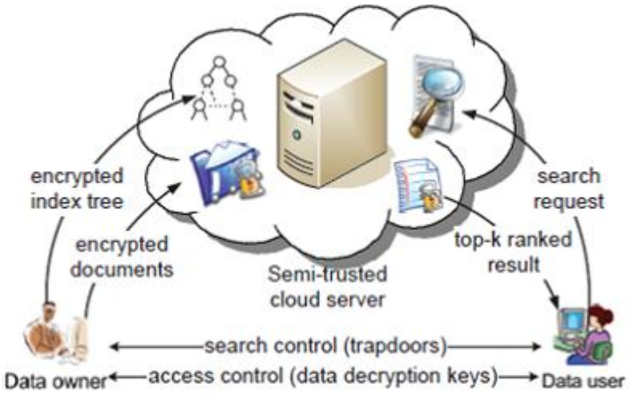

Fig -1: Architecture Diagram

There are three entities involved in the cloud computing paradigm –

1) Data Owner – It has collection of data documents that are to be stored onto cloud in an encrypted form. Encrypted index tree is also computed and stored in cloud.

2) Cloud Server – The job of cloud server is to store the encrypted documents of Data owner and provides search results in response of the user's search query.

3) Data User – These are the actual end users, who are querying to the cloud server with their search query and expects most relevant search results matching to the query.

When searching for a document for search keywords, an authorized user obtains a corresponding trapdoor through search control mechanism. After receiving trap door information from search user, cloud server searches in the Index and sends the matching documents matching the search query.

The cloud server is organized with a set of designated protocols which are to be strictly followed by both Cloud server owner & Data owners. These set of protocols clearly mentions the information that can be read by the server. The cloud server are "curious" but more "honest". Curios enough to understand and analyze the index and messages sent through the system. But also honest that it will not breach any protocols. There are two models –

1) Known Cipher text Model: Both the index and data documents are encrypted by the data owners.

2) Known Background Model: In this type of model, the cloud servers have more information than the previous one. It has trapdoor information and statistical information related to data documents.

3. Proposed System

To allow effectual similarity search, data owners constructs a secure index and subcontracts it to the cloud server. The data items are encrypted. Cloud servers accepts the search query and returns the search results without knowing any other information that the data owners has not shared. Multiple keyword rated search, results in rated similar results in response to the search query, instead of undistinguishable results.

The cloud servers are also not permitted to read any other information about the search keywords, the data documents and Index. It must strictly meet the privacy clause mentioned in client and server agreements. Overall bandwidth utilization should be minimum and thus resulting in faster response to the search query. This will improve the overall efficiency of the proposed system.

3.1 Introductory on Coordinate Matching

Coordinate matching [9] is a transitional similarity measure, it uses the number of search query keywords that matches the document to calculate the significance of that document in the search query. Boolean queries accomplishes very well results with the accurate search requirements of the user when exact subset of the data set to be regained are known. In practice this is not practical when huge amount of outsourced data is given. Hence, users provide list of interested keywords and fetches most significant documents with rated order.

3.2 Secrecy requirements of MRSE

Many traditional symmetric key cryptography algorithms can be used by the data owners, various operations on data documents are not described in this section. There are two main entities involved viz. Index and Search Query.

Following are four major steps involved in the algorithm –

1) Setup: Data owners generates a symmetric key and takes security parameter.

2) Build Index: Data owners generates Index based on data documents. This index is searchable in nature and is in encrypted form.

3) Trapdoor: This step generates trapdoor using search query keywords.

4) Query: Cloud servers accepts a query request and executes the rated search on the index after taking help of trapdoor.

The servers should acquire no information but search results in the searchable encryption. This is the typical secrecy guarantee in the related literature. With this general confidentiality description, we identified a bunch of strict privacy requirements explicitly for the MRSE framework.

3.2.1 Data Privacy

As for the data privacy, the data owner can recourse to the conventional symmetric key cryptography algorithms. This will encrypt the data before outsourcing, and also successfully avoid the cloud server from peeping into the outsourced data.

3.2.2 Index Privacy

If any information related to Index is known to the cloud server, it may learn the subject of the document. This is possible even cloud server gathers any association among encrypted documents from Index and searched keywords. If the data document is small, this index learning can lead to complete document data compromise [2]. Hence, proper care should be taken while constructing searchable index to avoid any kind of such attacks.

3.2.2 Search Privacy

Although above two privacy guarantees are demanded by default in the related literature, several search privacy requirements are involved in the query procedure. These are more complex and challenging to implement.

Keyword Privacy: It is important for users as they always wants to keep their search limited to them. The important thing is to hide the keywords entered by search user as to hide what users are searching. As this keyword related information can be utilized using reverse-engineering method and after performing some statistical calculation and an estimate can be made by the cloud servers to identify the keywords and document contents.

Trapdoor unlinkability: The fundamental protection for trapdoor unlinkability is to bring together sufficient non-determinacy into the trapdoor formation process so that the cloud server should not be able to realize the relationship of any given trapdoors, for example, to decide whether the two trapdoors are constructed by the single search request. In short – the trapdoor generation system should always calculate random values instead of being deterministic in nature.

Access Pattern: In the rated search, the access pattern is the sequence of search results where every search result is a set of documents with rank order. It is the results of frequent searches, and analytics can be done to form access pattern of specific search user based on this ranked search results.

4. Modules

Our anticipated system consists of following modules –

1) Data Owner Module

2) Data User Module

3) Encryption Module

4) Rank search Module

1) Data Owner Module:

Data owners are nothing but the actual owner of non-encrypted data documents. This module has two components. First user profile maintenance, which allows user to update his profile with login credentials. User registration is also done using this module. Second, user can encrypt his file with RSA algorithm and that file can be uploaded to cloud server. As the file is encrypted, unauthorized access to the file will be barred.

2) Data User Module:

Data users are the end users who requests access to the data stored onto cloud. This module has feature to register the users. This module is used to help the client to search the file by mentioning multiple keywords concept and get the exact result list based on the user's search query. The user is will select the required file and register the user details and get activation code in his mailbox. After inserting the action code ZIP file will be downloaded and user can download and extract the queried contents.

3) Encryption Module:

This module helps the cloud server to encrypt the data document using RSA Algorithm and the encrypted data document is converted to ZIP file with the activation code. This activation code is sent to the Data User via email.

File upload with encryption module and file download with encryption module are two primary functions of this module. File upload with encryption module allows the user to enter keywords that can be used to search the data document later in future. The data users can perform multiple keyword search on the cloud server. As the user finds out his requested file, the user asks for trapdoor file and the confirmation is took from the data owner for the same. IF the data owner approves the request, email is sent to the data user.

3) Rank Search Module:

This modules ensure the user to search the files that are searched frequently using rank search. This module allows the data user to download the file using his secret key to decrypt the downloaded data. This module allows the Data Owner to view the uploaded files and downloaded files.

3. CONCLUSIONS

The previous work mainly focused on single keyword search in encrypted cloud data documents. However intense need of searching multiple keywords to get accurate search results is been observed by many professionals. In this paper, we solve the same problem of multiple keyword rated search over encrypted cloud data, and found a variety of secrecy requirements. Amongst different multiple keyword semantics, we choose the effectual similarity measure of "coordinate matching," i.e., as many matches as possible, to effectively capture the relevance of outsourced documents to the search query keywords, and use "inner product similarity" to quantitatively assess such similarity measure. To support multiple keyword semantic without secrecy breaches, MRSE using secure inner product computation is proposed. Later, two improved MRSE schemes to achieve several strict privacy requirements in two different threat models are proposed. We also investigate some further enhancements of our rated search mechanism, including supporting more search semantics. Thorough analysis investigating privacy and efficiency guarantees of proposed schemes is given, and experiments on the real-world data set show our proposed schemes introduce low overhead on both computation and communication.

REFERENCES

[1] Ning Cao, C. Wang, Ming L., Kui Ren, W. Lou, "Privacy preserving Multi-Keyword Ranked Search over Encrypted Cloud data", IEEE Transactions, volume 25 No. 1, Jan 2014.

[2] Shiba Sampat Kale, "Privacy preserving multiple keyword ranked search with anonymous ID assignment over encrypted cloud data", International Journal of Computer Science and Information Technologies, Volume 5(6), 2014 ISSN : 0975-9646.

[3] Eduardo Fernández , Keiko Hashizume, David G R., "An analysis of security issues for cloud computing", Journal of Internet Services and Applications, 4:5, 2013.

[4] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data," Proc. IEEE INFOCOM, pp. 829-837, Apr, 2011.

[5] L.M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A Break in the Clouds: Towards a Cloud Definition," ACM SIGCOMM Comput. Commun. Rev., vol. 39, no. 1, pp. 50-55, 2009.

[6] P. Golle, J. Staddon, and B. Waters, "Secure Conjunctive Keyword Search over Encrypted Data," Proc. Applied Cryptography and Network Security, pp. 31-45, 2004.

[7] S. Kamara and K. Lauter, "Cryptographic Cloud Storage," Proc. 14th Int'l Conf. Financial Cryptograpy and Data Security, Jan. 2010.

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

[9] I.H. Witten, A. Moffat, and T.C. Bell, Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann Publishing, May 1999.

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

[11] E.-J. Goh, "Secure Indexes," Cryptology ePrint Archive, http:// eprint.iacr.org/2003/216. 2003.

[12] Y.-C. Chang and M. Mitzenmacher, "Privacy Preserving Keyword Searches on Remote Encrypted Data," Proc.

Third Int'l Conf. Applied Cryptography and Network Security, 2005.

[13] R. Curtmola, J.A. Garay, S. Kamara, and R. Ostrovsky, "Searchable Symmetric Encryption: Improved Definitions and Efficient Constructions," Proc. 13th ACM Conf. Computer and Comm. Security (CCS '06), 2006.

BIOGRAPHIES

Muzammil Ahmed has completed his B.E – Computer Science Engineering from MGM's College of Engineering. He is currently pursuing his M.E – Computer Science Engineering. His area of interests include Cloud Computing and SAP Administration etc.

Asrarullah Khan has completed his Masters. He is currently working as Assistant Professor in Matoshri Pratishthan Engineering college. His area of interests include Cloud Computing & Cryptography.