Techniques of keyword search over cloud data A Review

Vimmi Makkar
M.TechScholar
DCSA, MDU Rohtak

Sandeep Dalal
Assistant Professor
DCSA, MDURohtak

ABSTRACT
Cloud computing has become an integral part of IT industry. As it becomes prevalent, more and more sensitive information are being centralized in the cloud. A user stores his files in a cloud, and retrieves them whenever he wants to use them. This avoids the cost of maintaining as there is no need to personally install them on one's computer. But there is an extreme need for privacy of the data and the ability to search it without losing privacy. The users search their documents through keyword in plaintext. There is a cloud service provider (CSP), whose purpose is to provide more and more security and privacy. To keep user data confidential against an unfrosted CSP, an easy way is to apply cryptographic approaches, by disclosing the data decryption key only to authorized users. Thus the keyword privacy is maintained. In this paper, we define different types of keyword searching techniques.

Keywords
Cloud Computing, Ranked Keyword Search, security, and privacy preserving keyword search.

1. INTRODUCTION

Cloud Computing enables cloud customers to remotely store their data into the cloud so as to enjoy the on-demand high quality applications and services from a shared pool of computing resources [1]. Data can be stored in public or private cloud. The benefits brought by this new computing model for storage management, universal data access with independent geographical locations, and avoidance of capital expenditure on hardware, software, and personnel maintenances, etc [2].With the help of cloud computing, sensitive information can be centralized into cloud servers such as emails, personal records, private photographs and videos, company finance data and government documents.

One of the main concerns of user is to protect data privacy; that is why; sensitive data has to be encrypted before outsourcing so as to provide data confidentiality in cloud and beyond the cloud. Data encryption makes effective data utilization. In this we retrieve certain specific data files by keyword based search. This keyword search technique allows the users to retrieve the files of their interest. In this secure and privacy preserving keyword search on encrypted data approach contains three entities which are:

1. **User:** A user is an authorized person which stores the files, updates it, and control some functions like encryption and decryption.
2. **Key Server:** All the keys used for encrypting the document are stored by key server along with the signature of file names. After verifying the signature of file name the key server provides the key for encrypting a file to the user.
3. **CSP (Cloud Storage Server):** It is storage service providing data centre. The users store the cipher text of their file, the cipher text of the metadata of the files and the cipher text of the keyword into the cloud storage server, so that the server can know nothing about the information in the files and keywords.

2. RELATED WORK

In this section, main research areas related to keyword search are presented. When we talk about cloud service the work is more specific and the parametric. Many researchers performed a lot of work in the same direction.

In Year 2009, Georgia Koutrika presented a data cloud in which cloud search is performed on the basis of query summarization approach. He performed a query refinement model based on the summarization. Now based on this summarization the query is presented to the web architecture and relatively the search is performed for a reliable and effective cloud service [13].

A multimedia search for the cloud architecture is suggested by Wei-Ying Ma. In this work different multimedia services are suggested such as client PC, phone, TV etc. On the basic of the knowledge based search is performed to retrieve the multimedia analysis and will perform the search respective to the client request for the particular multimedia service [22]. Another tag based summarization approach is suggested by Byron Y.L. Kuof for the web search. The presented work is suggested on the public cloud. In which the integration of the web architecture and the database extraction is integrated. The work includes the refinement of the user query based on the cloud tags. The words extracted from the query are been summarized and this summarized query is passed to the public cloud. The cloud interface enabled the extraction of new and required information [20]. Another cloud search is suggested by Daniel E. Rose based on the information retrieval. The author presented his work on Amazon cloud service. The work is tested under different criteria such as scalability, configuration etc. The presented search reduce the barrier to allow a person or the organization to perform the content oriented search and the search is tested under the enterprises environment as well as on web search[21].

In Year 2012, Cengiz Orencik presented a rank based keyword search on the data cloud. In this work the document retrieval is performed on the cloud server based on the

keyword analysis and the information search is performed relative to the defined information. The presented work is performed on the encrypted data that has improved the security and the reliability of the retrieval. On this basis a secure protocol is suggested called Private Information Retrieval. The system will performed the query and present the final results on the basis of parametric ranking. The presented work is the efficient computation and communication of the requirement analysis [14].

Mathew J. Wilson performed a work based on web search engine based for the keyword cloud. In this work the clouds are represented by some tags called the Meta data. The Meta data defines the cloud with relative parameters in terms of the services will be done under different parameters. The first parameter considered here is the most appropriate of its security, efficiency and the reliability criteria. On the basis of this the keyword match is performed on different cloud keywords. The work includes the learning stage for the keyword extraction and the comparative analysis is performed to extract the related cloud services from the system [19].

3. PROPOSED TECHNIQUES FOR KEYWORD SEARCH

3.1 Secure Ranked Keyword Search over Cloud Data

Development of a private cloud is very expensive. Storage of sensitive data in public cloud is very risky. To make it possible, unauthorized access is avoided by storing the data in encrypted format. This paper tackles the problems of enabling searchable encryption system with support of secure ranked search in order to implement the top k retrieval. In this paper, statistical measure approach from IR and text mining to embed weight information of each file during establishment of searchable index before outsourcing the encrypted file collection is explored.

Team frequency: Number of times a particular keyword appears within the file.

Inverse document frequency (IDF): It is calculated as the total number of files by the number of files in particular keyword.

Ranking function: It is calculated by using TF*IDF rule.

Information leakage is avoided by using one too many order preserving mapping. The basic scheme, security of ranked searchable encryption is the same as previous SSE schemes, i.e. the users gets the ranked results without letting cloud server learn any additional information more than the access pattern and search pattern. The scheme clearly satisfies the security guarantee of SSE, i.e. only the access pattern and search pattern is leaked.however, ranking is done on the user side, which may bring in huge computation and post processing overhead. Moreover, sending back all the files consumes large undesirable bandwidth. In this way, server still learns nothing about the values of relevance scores, but it knows the requested files are more relevant than the unrequested ones, which inevitably leaks more information than the access pattern and search pattern [15]. For efficient use of outsourced cloud data, enabling ranked searchable encryption is the purpose of this approach. The main goals of this system are: I) Ranked keyword search ii) Security guaranteeiii) Efficiency.

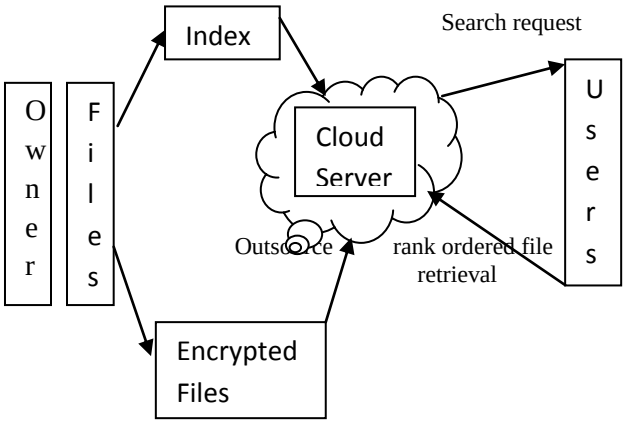

Fig 1: Architecture of the search over encrypted cloud data

Considering a cloud data hosting service with three different entities i.e. *data owner* (O), *data user* (U), and *cloud server* (CS). Data owner has a collection of *d*ata files $C = (F1, F2, FN)$ that he wants to outsource on the cloud server in encrypted form. for doing this ,before outsourcing, on cloud server data owner will first build a secure searchable index I from a set of m distinct keywords $W = (w1, w2, wm)$ extracted from the file collection C, and store both the index I and the encrypted file collection *C* on the cloud server. [15]

Considering that the authorization between the data owner and users is properly done. A certified user submits a keyword search request in a secret form —a trapdoor *Two* of the keyword—to the cloud server. Upon receiving the search request *Two*, the cloud server is responsible to search the index *I* and return the equivalent set of files to the user. The secure ranked keyword search problem is: result of the submitted search should be returned according to certain ranked relevance criteria (e.g., keyword frequency based scores), to improve file retrieval accuracy for users without previous knowledge on the file collection *C*.

The searching in the encrypted data is done by using ranking function and the retrieved results are authenticated. [15], [16]

3. 2 Authenticated and Complete Query Results from Cloud Storages

The query results can be considered only by evaluating two aspects: first is the authentication and the other is the completeness of the query results. The authentication of the retrieved data is done by verifying the corresponding digital signatures one after the. The goal can also be achieved by using authenticated data structures including signature aggregation [3], signature chaining [4], or hash tree structure [5]. The requirement of computation of an aggregated value by the users in these designs renders this computation intensive. Along with this there is no guarantee of completion of results, only authentication of returned data is ensured. The most applicable research is verifying the completeness of the query results from the database management system (DBMS)

[6], [7]. The problem with this is that only the dataset is considered in plaintext and there is requirement of searchable field to define some kind of ordering, which is hard to achieve when the data is encrypted. A new cryptographic primitive is introduced called Order Preserving Encryption (OPE) [8], which reserves the order of the data after the data is encrypted. Therefore, one original authentication scheme for the authenticity and completeness of the query results are required for all of the cloud users. The three threats to query replies under this model are:

1) bilinear paring 2) searchable encryption 3) message authentication code .In this one scheme for CSPs is proposed to the proof of a query results for the cloud users and then it is demonstrated that this scheme uniquely achieves the authentication and completeness of the query results from the CSPs. Secondly, conduction of an extensive computation and communication analysis is performed. This makes the scheme both practical and feasible for the cloud users.

The system models consist of entities.1) Cloud Storage Client (CSC) stores a large number of data in cloud storage server. A CSC could be an individual user or an organization. 2) Cloud Storage Provider (CSP) provides search-based secure retrieval services for cloud storage client. A CSP has rich storage space and computation power to manage the data of CSCs. 3) Trusted Authority (TA) is trusted by all the other system entities, and issues identification to them. [**9**]

The system model consists of 3 system entities: **CSC** encrypts the data and generates MAC .**CSP**; here data corresponding to MAC is stored.

The proposed scheme ensures secure and efficient authentication to assure the security of query results. It takes advantages of the PEKS scheme to construct searchable indexes. The design ensures that the encrypted file indeed contains certain keywords without decrypting the files. There is introduction of a counter for each keyword to assure the completeness of query results. Because no unqualified files can be included without being caught and the number of returned files is correct, the CSC can be assured of the correctness of the query results.

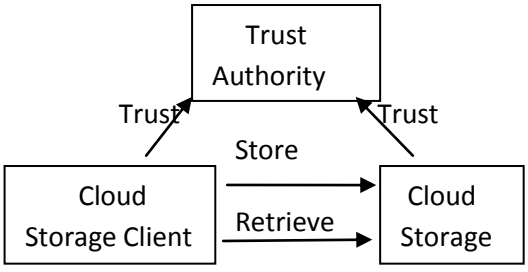

Fig 2: Cloud Storage Access Model

3.3 Privacy Preserving Encrypted Keyword Search

Qin liu introduced a privacy preserving keyword search scheme in cloud computing. A secure and efficient privacy preserving encrypted keyword search scheme suitable for cloud storage is proposed. In this secure and privacy preserving search keyword on encrypted data approach contains three participants are there: 1) user 2) key server 3) CSP.

CSP contains many schemas that use the system models like- Public key encryption with keyword searching (PKES), Efficient and Privacy Preserving Keyword Search (EPPKS) and Secure and Privacy Preserving Keyword Search (SPKS). These types of the schemas generally use the following function in the given order:

- Key Generation
- Email Encryption
- Keyword Encryption
- Trapdoor Computing
- Testing
- Decryption/ Partial Decryption
- Recovery

The goal of this schema is: The document stored in cloud storage is not accessed by an unauthorized user. It also used to avoid the statistical attack on the cloud storage. [10],[11], [12]

3.4 Multi-user Private Keyword Search

Almost all of the searchable encryption schemes which are existed that work in the single-user setting: only can issue valid search queries. Observing that in enterprise-outsourcing-database, there is often need of a multi-user searchable encryption, a multi user private keyword search for cloud computing is developed. In multi user setting some more factors are to be considered than in the single- user setting, that are user accountability, user dynamics (it means the joining of new users and revocation of existing users). We can achieve multi-user searchable encryption by directly applying a single-user scheme to the multi-user setting by sharing secret query key among the group of multiple users. Curtmola et al. proposed a very different approach for achieving multi-user searchable encryption, in which there is transfer of single-user searchable encryption scheme to working in the multi-user setting. In this users encrypt their search queries using the broadcast encryption before submitting to the server who hosts the database. Since the server also knows the broadcast encryption key, it thus can decrypt and obtain user search queries. User dynamics guarantees that only the set of users which are authorized and also the server can use broadcast encryption. This method ensures that a user who is not authorized cannot use valid search results to the server and the revoked users still retain their ability to search as they can search encrypted database as long as they are given database. But it is expensive and using this encryption scheme cannot achieve user accountability. According to the Application Scenario of enterprise outsources its database to the cloud, the system consists of $\{D, \text{ENT}, \text{CLD}, \mathcal{U}\}$, where D is a database, ENT is the enterprise, CLD is the cloud providing storage services, and $\mathcal{U}$ is a set of users. The database D is composed of a number of records $\{d1, d2, \cdots\}$ of multiple attributes, and one attribute is keyword used for search. Each authorized user $u \in \mathcal{U}$ is issued a distinct query key qku by ENT, and can issue search queries based on her chosen keywords by using qku and then CLD is expected to return reply of the query.

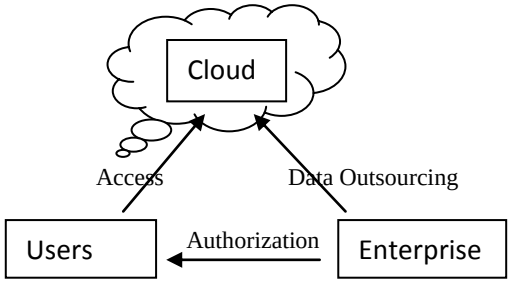

Fig 3: Application scenario of enterprise-outsourcing-database to cloud

3.5 Fuzzy Keyword Search

This keyword search deeply enhances system usability by returning the matching files when users' searching inputs accurately match the predefined keywords or the closest possible matching files based on keyword similarity semantics, when exact match fails. Usage of edit distance to quantify keywords similarity and development of a technique for construction of fuzzy keyword sets. Fuzzy eliminates the need for enumerating all the fuzzy keywords and the resulted size of fuzzy keywords sets is extensively concentrated. Goals of introducing fuzzy keyword search are: I) to discover new mechanism for constructing storage efficient fuzzy keyword sets; ii) to design well-organized and effective fuzzy search scheme based on the constructed fuzzy keyword sets; iii) to validate the security of the planned scheme.[17], [18]

Main modules in Fuzzy keyword search are:

a) Wildcard-based technique: To edit the operations at the same position a wild card based technique is used. We can calculate the edit distance by using substitution, deletion and insertion.

b) Gram-based technique: Here the fuzzy set is constructed based on grams. The gram of a string is a substring and it can be used for efficient estimated search. The order of the characters after the primitive operation is always kept the same before the operations.

c) Symbol-based tire-traversed scheme: In this technique, for storing the fuzzy keyword set over a finite symbol set, a multi way tree is constructed. Here we consider a cloud data system consisting of data owner, data user and cloud server. Given her a collection of *n* Encrypted data files *C* = (F1, F2, FN) stored in the cloud server, a predefined set of distinct keywords W = (w1, w2, we), the cloud server provides the search service for the authorized users over the encrypted data *C*. We assume that the authorization between the data owner and users is appropriately done. An authorized user types in a request to selectively retrieve data files of user's interest. The cloud server is responsible for mapping the searching request to a set of data files, where each file is indexed by a file ID and linked to a set of keywords. The fuzzy keyword search scheme returns the search results according to the following rules:

If the user's searching input exactly matches the pre- set keyword, the server is estimated to return the files containing the keyword; if there are typos and/or format inconsistencies in the searching input, the server will return the closest possible results based on pre-specified similarity semantics.

4. CONCLUSION

As we know that cloud computing is the latest innovative technology. In this, a user can store his personal and private files in a cloud and can retrieve them whenever he wants them. In this paper we reviewed different techniques for keyword search. As our ongoing work, we will continue to research on ranking criteria which is specified on the basis of implicit and explicit properties such as content relevance, security, availability etc.

5. REFERENCES

[1] P. Mell and T. Grace, "Draft nits working dentition of cloud com- putting," Referenced on Jan.23rd,2010online http://csrc.nist.gov/groups/SNS/cloud-computing/index.html, 2010.
[2] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Kong- win ski, G. Lee, D. A. Patterson, A. Rabin, I. Stoical, and M. Zaharias, "Above the clouds: A Berkeley view of cloud computing," University of California, Berkeley, Tech. Rep. UCB-EECS-2009-28, Feb 2009
[3] E.Mykletun,M.Narasimha, and G.Tsudik, "Authentication and intesgrity in outsourced databases," Trans. Storage, vol. 2, pp. 107–138, May 2006.
[4] M. Narasimha and G. Tsudik, "Dsac: integrity for out-sourced databases with signature aggregation and chaining," in Proceedings of the 14th ACM international conference on Information and knowledge management, ser. CIKM '05. New York, NY, USA: ACM, 2005, pp. 235–236.
[5] H. Pang and K. Mouratidis, "Authenticating the query results of text search engines," Proc. VLDB Endow., vol. 1, pp. 126– 137, August 2008.
[6] M. Narasimha and G. Tsudik, "Authentication ooutsourced databases using signature aggregation and chaining," in Database Systems for Advanced Applications, ser. LNCS. Springer, 2006, vol. 3882, pp. 420–436.
[7] H. Pang and K.-L. Tan, "Verifying completeness of relational query answers from online servers," ACM [8]
[8] Trans. Inf. Syst. Secure., vol. 11, pp. 5:1–5:50, May 2008.Boldyreva, N. Chenette, and A. O'eill, "Order-preserving encryption revisited: Improved security analysis and alternassss- tive solutions," vol. 6841, pp. 578–595, 2011.
[9] Fu-Kuo Tseng and Yung-Hsiang Liu ,Rong-Jaye Chen: Toward Authenticated and Complete Query Results from Cloud Storages,appear in IEEE publication.
[10] Liu Hong-xia, Dai Jia-zhu, Jiang Chao: Research on Privacy Preserving Keyword Search in Cloud Storage, appear in IEEE publication, 978-1-4244-5540-9/10, 2010.
[11]Qin Liuy, Guojun Wang, and Jie Wuz: An Efficient Privacy Preserving Keyword Search Scheme in Cloud Computing.
[12] Qin Liuy, Guojun Wang, and Jie Wuz: Secure and privacy preserving keyword searching for cloud storage services, appear in Journal of Network and Computer Applications, 9 March 2011.
[13] Georgia Koutrika (2009),"Data Clouds: Summarizing Keyword Search Results over Structured Data", EDBT 2009, March 24–26, 2009, Saint Petersburg, Russia. ACM, p 391-402
[14] 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

[15] Cong Wang, Ning Cao, Kui Ren and Wenjing Lou: Secure Ranked Keyword Search over Outsourced Cloud Data, IEEE publication, 2010

[16] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data,Proc. IEEE INFOCOM '11,2011
[17] Jin Li, Qian Wang, Cong Wang, Ning Cao , Kui Ren, and Wenjing Lou: Fuzzy Keyword Search over Encrypted Data in Cloud Computing, IEEE publication, 2010
.[18] S.T.Balamuralikrishna, C.Anuradhaand: Fuzzy keyword search over encrypted data over cloud computing ,Asian Journal of Computer Science and Information Technology2011.
[19] Mathew J. Wilson (2012)," Keyword Clouds: Having Very Little Effect on Sensemaking in Web Search Engines", CHI 2012, May 5–10, 2012, Austin, Texas, USA,ACM,p2069-2074
[20] Byron Y-L. Kuo (2007)," Tag Clouds for Summarizing Web Search Results", WWW 2007, May 8–12, 2007, Banff, Alberta,Canada.pp.1203
[21] Daniel E. Rose (2012)," CloudSearch and the Democratization of Information Retrieval" Wei-Ying Ma (2009)," Rethinking Multimedia Search in the "Clients + Cloud" Era", LS-MMRM'09, October 23, 2009, Beijing, China. Pp. 1-1