Assuring Integrity in Privacy Preserving Multi-keyword Ranked Search over Encrypted Cloud Data

CH. Keerthi prathyusha
Department of Computer Science and Engineering, JNTUA College of Engineering, Anantapuramu, A.P, India.

Dr. C. Shoba bindhu
Associate Professor, Department of Computer Science and Engineering, JNTUA College of Engineering, Anantapuramu, A.P, India.

Abstract – Cloud computing is a general term for anything that involves delivering hosted services over the Internet and centralize their sensitive data into the cloud. With a mass of data files stored in the cloud server, it is important to provide keyword based search service to data user. However, in order to protect the data privacy, sensitive data is usually encrypted before outsourced to the cloud server, which makes the search technologies on plaintext unusable. In this paper, we propose a multi-keyword ranked search plan over the encrypted cloud data, which at once meets a set of strict privacy requirements. We choose "coordinate matching" so that it gives as many as possible matches to the search query. When users given a search query it gives all the relevant data based on the keyword and it provides ranking based on the search result. The proposed scheme not only searches the exact files but also it get back the related results based on the query keyword.

Index Terms – Sensitive data; multi-keyword ranked search; searchable encryption, Ranked search.

1. INTRODUCTION

Due to the quick development of data, the data owners want to save their data information into the cloud to release the trouble of data storage and protection [1]. But, the cloud customers and the server are not in the same trustworthy domain, the outsourced data may cause risk while display. The sensitive data has to be encrypted before out sourcing and protect the data. The traditional information retrieval (IR) has previously allowed multi-keyword ranked search for the data user. The cloud server has to provide the information to the data user with the related function, secure data and search privacy, when it stores the data in to cloud server and it provides easy searched and utilized.

In the literature, searchable encryption techniques [2-4] are providing secure search to the data users with encrypted data. Based on the set of files it build a searchable inverted index that stores the files according to the list of mapping keywords. When data user's searches for a keyword, a trapdoor is generated for this keyword and then submitted to the cloud server. When it receives the trapdoor the cloud server allow comparison between the trapdoor and index, and then it returns all gathered information to the data users based on their search query. But, these methods only allow exact single keyword search. Some researchers study the problem on secure and ranked search over outsourced cloud data. Wang *et al.*, [10] propose a secure ranked keyword search scheme. Their solution combines inverted index with order-preserving symmetric encryption (OPSE). In terms of ranked search, the order of retrieved files is determined by numerical relevance scores, which can be calculated by $TF \times IDF$. The relevance score is encrypted by OPSE to ensure security. It enhances system usability and saves communication overhead. This solution only supports single keyword ranked search. Cao *et al.*, [6] propose a method that adopts similarity measure of "coordinate matching" to capture the relevance of files to the query. They use "inner product similarity" to measure the score of each file. This solution supports exact multi-keyword ranked search. It is practical, and the search is flexible. Sun *et al.*, [7] proposed a MDB-tree based scheme which supports ranked multi-keyword search. This scheme is very efficient, but the higher efficiency will lead to lower precision of the search results. In cloud computing, data owners may allocate their outsourced data information with many users, according to their search keyword it gives only retrieve data files [3] it is better way to use keyword–based retrieval and this is in form of scenarios. To improve more feasibility in cloud it gives all the relevant files that match to the search keyword and then it should be ranked based on their search query and then High relevance files are sent back to users. Searchable symmetric encryption (SSE) allows enable search on cipher text these traditional SSE scheme provide secure to users while retrieving but it supports only Boolean keyword search. In Searchable Symmetric Encryption (SSE) to improve more efficiency and secure it supports top-k single keyword retrieval.

In MRSE, for protecting the data privacy and security data encryption may help some extent encrypted data has to retrieve by the cloud through Searchable Symmetric Encryption (SSE) to eliminate the leakage of data privacy we focus on SSE in these we are considering the aspects of similarity relevance and scheme robustness. Where the server side ranking is based on order-preserving encryption (OPE) it may leaks the data

privacy so to eliminate the loss of data privacy we propose a Multi-Keyword Ranked Searchable Encryption and it supports the top-k Multi-keyword retrieval. In MRSE, it allows a Vector space model and Homomorphic encryption. Where the vector space provides sufficient search accuracy based on search query and Homomorphic allow users to involve in the ranking based on the majority of computing work on server side by operations.

In addition, fuzzy keyword search have been developed. These methods employ a spell-check mechanism, such as, search for "wireless" instead of "wireless", or the data set-up may not be the same *e.g:* "data-mining" versus "data mining". Chuah *et al.*, [8] propose a privacy-aware bed-tree method to support fuzzy multi-keyword search. This approach uses edit distance to build fuzzy keyword sets. Bloom filters are constructed for every keyword. Then, it constructs the index tree for all files where each leaf node a hash value of a keyword. Li *et al.*, [9] develop change distance to checking keywords similarity and construct storage-efficient fuzzy keyword sets. Specially, the wildcard-based fuzzy set construction approach is designed to save storage overhead. Wang *et al.*, [5] employ wildcard-based fuzzy set to build a private trie-traverse searching index. In the searching phase, if the edit distance between retrieval keywords and ones from the fuzzy sets is less than a predetermined set value, it is considered similar and returns the corresponding files. These fuzzy search methods support tolerance of minor types and inconsistencies, but do not support semantic fuzzy search. Considering the existence of polysemy and synonymy [11], the model that supports multi-keyword ranked search and semantic search is more reasonable.

2. PROBLEM FORMULATION

2.1. System Model

The system model can be considered as three entities, as depicted in Figure 1 as follows:

Data owner has a set of data documents D {*d* 1 , *d* 2 , ..., *d m* } . A set of distinct keywords W = {w 1, *w* 2,...,*w n*} is extracted from the data collection *D* . The data owner will firstly construct an encrypted searchable index *I* from the data collection *D*. All files in *D* are encrypted and form a new file collection, *C*. Then, the data owner uploads both the encrypted index and the encrypted data collection of the cloud server.

Data user provides keyword to the cloud server search control mechanism is generated through trapdoor. In this paper, we assume that the authorization between the data owner and the data user is approximately done.

Cloud server received T_w from the authorized user. Then, the cloud server calculates and returns to the corresponding set of encrypted documents. Moreover, to decrease the computational cost and time where the data user returns a k values along with the trapdoor *t* key to the cloud server so then cloud sends only top-k files that are mostly matches to the search query.

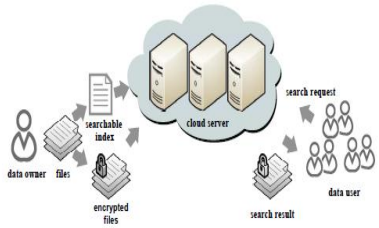

Figure 1**:** Architecture of Ranked Search over Encrypted Cloud Data

In this paper, we will solve the problem of multi-keyword ranked search and retrieve the most relevant files. We define multi-keyword ranked search which supports top-k retrieval multi-keyword ranked search. By using the proposed scheme shows not only the correct matching files but also related searches of the keyword query. For example, when the user search for the keyword "automobile" to search files, the proposed method searches the related contents files for the keyword "automobile", but also shows the related documents search including the term "car" We take a large matrix of term-document association data and construct a semantic space wherein terms and documents are closely associated are placed near one another. For meeting the confront of supporting multi-keyword ranking without privacy breaches, we propose the basic idea: the multi-keyword ranked search (MRSE).

The paper is organized as follows. In Section, we describe the system model, privacy requirements, and notations. Section provides the detailed description of our proposed mechanism. Section presents the experiment and security analysis. Section summarizes the conclusion.

2.2. Threat models and Design Goals

The cloud server is considered as "honest-but-curious" in our model. Particularly, the cloud server both follows the designated protocol specification but at the same time it analyzes the data in its store and received the message flows during the additional information has to gather from protocol [12]. The drawback in these paper focus on the basic modelling of the multi-keyword ranked search which consumes more computational time and also degraded performance It provides the efficient results for given search query after going through ranking process assigned to the search results obtained but it may not provide data privacy for the ranking method efficiently.

In this paper, we purpose to achieve security and ranked search under the above model. The designed goals of our system are following:

- Multi-keyword Ranked Search**:** It supports both multi- keyword query and support result ranking.
- Privacy-Preserving**:** Our scheme is having designed to meet the privacy requirement and prevent the cloud server from getting additional information from index and trapdoor and providing privacy for the search keyword.
- *Index Confidentiality*. The *TF* values of keywords are stored in the index. Thus, the index stored in the cloud server needs to be encrypted.
- *Trapdoor Unlinkability*. The cloud server performs some statistical analysis to the search result. Meanwhile, the same query should generate different trapdoors when searched twice. The cloud server is acting like intermediate between the search query information and it should not realize the relationship between trapdoors. Here the trapdoor is the key to the secret information to provide privacy for the search query.

2.3. Coordinate Matching

In MRSE, we choose "coordinate matching" to provide the relevant results for the query. When the data user sends the search query to the data owner then it searches all the documents in the dataset and it return back as matches as possible results to the data user. It provides more computational time and it performs time consuming for searching the data.

3. THE PROPOSED SCHEME

In proposed scheme we solve the problem of Multikeyword Ranked search over encrypted cloud data setup a privacy requirements for secure utilization of cloud data where we propose a basic model of MRSE to remove the leakage of data privacy. In the majority part of computing work is done on cloud and the users takes part in ranking where top-k multikeyword retrieval over encrypted cloud guarantees the high security and efficiency we suggest two MRSE schemes on "coordinate matching" where it gives as many as possibilities of search query and it also allows the concepts of similarity with relevance and scheme robustness [13]. We allow top-k retrieval over encrypted cloud for accurate ranking and integrity for the search results. In MRSE, proposed system will investigate checking the integrity of rank order in search results [12] where it supports the multi keyword top-k retrieval under encrypted cloud data.

KeyGen is key generation algorithm that is run by data owner and where it compute and upload the file into the cloud and it having an Index construction where it works as an independent process the keywords are constructed in an index with encrypted data over cloud. The key is providing to the index encrypted by data owner. In server side we provide encrypted top-k retrieval based on ranking with no of occurrence of keyword then it provide top-k searches based on the query keyword. Then the data owner provide secret key to the user so that the file decrypt and download it where if the match results are similar then the signature verification is passed.

System architecture is the conceptual design that defines the structure and behavior of a system.

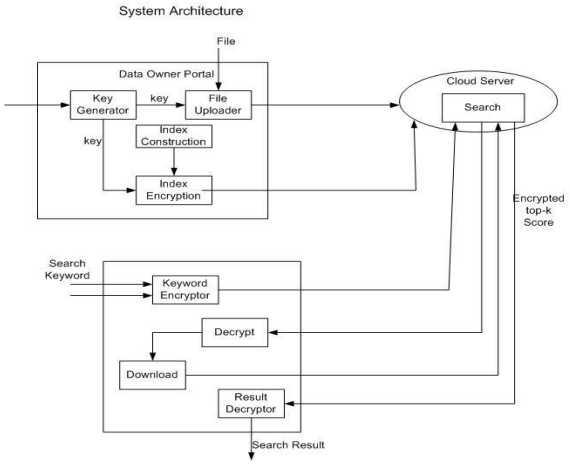

Figure 2**:** System Architecture

4. PERFORMANCE AND SECURITY ANALYSIS

In this section, we show a through experimental estimation of the proposed technique on a real dataset. The whole experiment is implemented by C language on a computer with Core 2.83GHz Processor, on Windows7 system. For the proposed scheme, we will reduce to separate dimensions and results provide accurate and related documents for the search query. When the cloud server get backs the top-k relevant documents based on the data vector and converts to query vector then it shows the frequently occurred top-k keyword documents and it were received by the user, the performance of our method is compared with the original MRSE scheme.

4.1. Efficiency

The proposed scheme is depicted in details in previous section, except the KeyGen algorithm. In our scheme, we adopt Gauss-Jordan to compute the inverse matrix. The time of generating key is decided by the scale of the matrix. Besides, the proposed scheme that processed by *SVD* algorithm will consume time. Other algorithms, such as index construction, trapdoor generation, query, which is put forward by us, are consistent with the original MRSE in time-consuming.

Index construction*:* In constructing index first we have to build a searchable sub-index. The first step is to extracted documents and map according to their keyword set and also data vector has to be in encrypted documents. Fig 3 shows that the time cost of encrypting and mapping is depending on the data vector and also its dimensionality of data vector. Based on the documents in the data set it constructs the whole index it is also

relate to the sub-index which is similar to the number of keywords in the dataset through it calculates the time cost of building the index.

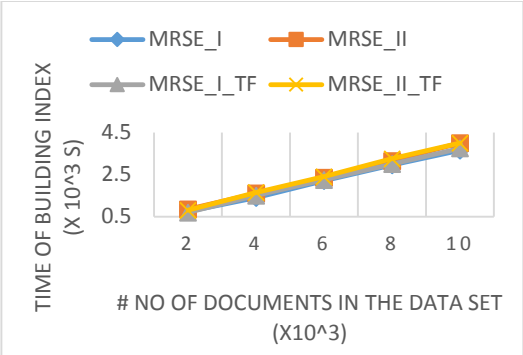

Figure 3: Time cost of building index

Query*:* query execution in the cloud server considering the ranking and computing for all the documents in the data set. The query time is calculated by the number of documents in the data set. If the documents have similar query keywords in the dataset it retrieves all the information quick based on the dimensionality of data vector and query vector.

4.2. F-measure

In this paper, we still use the measure of traditional information retrieval. Before the introduction of the F-measure's concept, we will firstly give the brief of the precision and recall. Precision is the fraction of retrieved the accurate relevant documents while recall is related fraction that are retrieved. Both the recall and precision are based on the occurrence and relevance F-measure that combines both precision and recall in harmonic mean. Here, we adopt F-measure to weigh the result of our experiments.

Precision: In each data vector dummy keywords are inserted and then it selected in every query. While retrieving it does not shows the accurate results based on the dummy keywords. When the cloud server get back top-k documents the original similarity is decreased because of inserting dummy keywords in data vector it shows some documents of unwanted top-k to the user. For supply secure and privacy top-k documents to the user. We define as precision to get real top-k documents return by server.

Ranking*:* In cloud server the user's rank privacy may leak in small consequences to provide privacy guarantee firstly the number of documents should be ranked when it retrieve top-k documents. It provides rank privacy for every document. Sometimes it shows higher precision but low ranking privacy assurance then it provide high ranking privacy and low precision our scheme balance both parameter to satisfy data users' needs on precision and ranking.

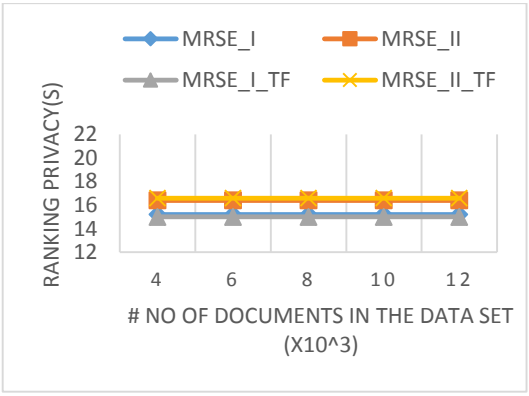

Figure 4: Ranking Privacy

4.3. Performance Analysis

For a clear comparison, our proposed scheme gets score higher than the original *MRSE* in F-measure. Since the original scheme employs exact match, it must miss some similar words which is similar with the keywords. However, our scheme can make up for this disadvantage, and retrieve the most relevant files.

- *Trapdoor Unlinkability*. The trapdoor of query vector is cause from a random splitting operation, which it gives the same search requests are change into different search trapdoors. And thus, the query unlinkability is well preserved.
- *Keyword Privacy*: In the known background scheme, the cloud server is supposed to have more information, such as the distribution *TF* values of keywords in the dataset. The cloud server is able to identify keywords by analyzing these specific distributions. In our scheme, the *TF* distributions of keywords will be leaked directly when there is only one query keyword. Thus, our proposed scheme considered to obscure the *TF* distributions of keywords with the duplicate values. That is to say the keyword privacy is protected.

5. CONCLUSION

In this paper, a multi-keyword ranked search data is proposed, which meanwhile supports top-k retrieval ranked method. We use "coordinate matching" technique these helps to find out as many as possible matches of the search query however security and privacy are major barriers for the users to get used to cloud so to provide integrity for the ranking search results we allow MRSE. The effective search of the data which is encrypted in a cloud is extracted by the search query which will be encrypted at the client end. The proposed system not only gives the exact result but it also find out all the relevant data based on the search keyword. In server side ranking supports Order-Preserving Encryption (OPE) it may cause some leakage of data privacy so, to eliminate the leakage we propose the top-k multikeyword retrieval so that, we can eliminate leakage of

data privacy and provide integrity for ranking in the search result using coordinate matching while assuming that the cloud server is untrusted.

As our future work, we will concentrate on the encrypted data of ranking keyword search in order that we can confront with the more sophisticated search.

REFERENCES

[1] 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.
[2] S. Kamara and K. Lauter, "Cryptographic Cloud Storage," Proc. 14th Int'l Conf. Financial Cryptograpy and Data Security, Jan. 2010.
[3] N. Cao, S. Yu, Z. Yang, W. Lou, and Y. Hou, "LT Codes-Based Secure and Reliable Cloud Storage Service," Proc. IEEE INFOCOM, pp. 693-701, 2012.
[4] 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.
[5] 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.
[6] I.H. Witten, A. Moffat, and T.C. Bell, Managing Gigabytes:Compressing and Indexing Documents and Images. Morgan Kaufmann Publishing, May 1999.
[7] D. Song, D. Wagner, and A. Perrig, "Practical Techniques for Searches on Encrypted Data," Proc. IEEE Symp. Security and Privacy, 2000.
[8] E.-J. Goh, "Secure Indexes," Cryptology ePrint Archive, http:// eprint.iacr.org/2003/216. 2003.
[9] 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.
[10] A. Singhal, "Modern Information Retrieval: A Brief Overview,"IEEE Data Eng. Bull., vol. 24, no. 4, pp. 35-43, Mar. 2001.
[11] D. Boneh, G.D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public Key Encryption with Keyword Search," Proc. Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT), 2004.
[12] M. Bellare, A. Boldyreva, and A. ONeill, "Deterministic and Efficiently Searchable Encryption," Proc. 27th Ann. Int'l Cryptology Conf. Advances in Cryptology (CRYPTO '07), 2007.
[13] M. Abdalla, M. Bellare, D. Catalano, E. Kiltz, T. Kohno, T. Lange, J. Malone-Lee, G. Neven, P. Paillier, and H. Shi, "Searchable Encryption Revisited: Consistency Properties, Relation to Anonymous Ibe, and Extensions," J. Cryptology, vol. 21, no. 3, pp. 350-391, 2008.
[14] P. Golle, J. Staddon, and B. Waters, "Secure Conjunctive Keyword Search over Encrypted Data," Proc. Applied Cryptography and Network Security, pp.31-45, 2004.
[15] D. Boneh and B. Waters, "Conjunctive, Subset, and Range Querieson Encrypted Data," Proc. Fourth Conf. Theory Cryptography (TCC),pp. 535-554, 2007.

Authors

Ch.Keerthi Prathyusha received B.Tech degree in Computer Science and Engineering from Rajeev Gandhi Memorial College of Engineering and technology, Nandyal affiliated to JNTUA College of Engineering, Anantapuramu, A.P, India, during 2009 to 2013. Currently pursuing M.Tech in Software Engineering from JNTUA College of Engineering, Anantapuramu, A.P, India, during 2013 to 2015 batch. Her Area of interests includes Cloud Computing, Software Testing.

C. Shoba Bindu is an Associate Professor of Computer Science and Engineering at Jawaharlal Nehru Technological University College of Engineering, Ananthapuramu. She obtained her Bachelor degree in Electronics and Communication Engineering, Master of Technology in Computer Science from Jawaharlal Nehru Technological University Hyderabad and Ph.D. in Computer Science and Engineering from Jawaharlal Nehru Technological University Anantapuramu. She has published several Research papers in National International Conferences and Journals. Her research interests includes network security and Wireless communication systems.