ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Biology, Engineering, Science and Technology (IJARBEST)

Vol. 1, Special Issue 5, May 2013

SECURE AND EFFICIENT MULTIKEYWORD TOP-K RETRIEVAL OVER ENCRYPTED CLOUD DATA

Mrs. D. Dhanya, Assistant Professor
Mar-Ephraem College of Engineering & Technology
dhanvis@gmail.com

ABSTRACT- Cloud computing is a rapidly growing technique for storing large volume of data. The characteristics of cloud like on-demand self-service and measured service provide scalable, secure, reliable and cost effective services to the cloud data users. As cloud computing become prevalent, sensitive information's are being increasingly centralized into the cloud. In order to ensure the protection of data privacy, sensitive data has to be encrypted before outsourcing which makes the effective data utilization very difficult. An encrypted cloud data search for the effective data retrieval over large number of data users and documents is essential. Keyword search techniques are used for this purpose. A new scheme called Two Round Searchable Encryption (TRSE) scheme with fuzzy keyword search is proposed to eliminate the data privacy leakage arises when using traditional encryption schemes over the encrypted cloud data for retrieval. The proposed TRSE scheme with fuzzy keyword search supports multikeyword top-k retrieval greatly enhances system usability by returning the matching files when users searching inputs exactly matches the predefined keywords or the closest possible matching files based on keyword similarity semantics. To Perform TRSE with fuzzy keyword search, wild-card based technique is used. This technique quantifies keywords similarity and develops an advanced technique on constructing fuzzy keyword sets, which greatly reduces the storage and representation overheads.

Keywords – Two Round Searchable Encryption, Multikeyword, Fuzzy, Wild card technique

1. INTRODUCTION

Cloud computing is a technology that uses the internet and central remote servers to maintain data and applications. Cloud computing is the delivery of computing and storage capacity as a service to a community of end recipients. Cloud computing presents a new way to supplement the current consumption and delivery model for IT services based on the Internet, by providing dynamically

ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Biology, Engineering, Science and Technology (IJARBEST)

Vol. 1, Special Issue 5, May 2013

scalable and often virtualized resources as a service over the Internet**.** The data processed on clouds are often outsourced, leading to a number of issues related to accountability, including the handling of personally identifiable information and revealing of user's confidential data. Consider the situation that, users need to be able to ensure that their data are handled according to the service level agreements made at the time they sign on for services in the cloud. So it became essential to provide an effective mechanism for users to ensure that their confidential data are protected from neither reaching unauthorized persons nor cloud server learning. Conventional mechanisms for outsourcing the data were done in plain text manner. It is very difficult to retrieve appropriate data from the large cloud data collection. Cloud computing supports keyword based retrieval to only retrieve the data files they are interested in.

To protect the privacy of the data to be outsourced, data files should be encrypted before storing it into the cloud server. This makes effective data utilization very difficult. Several searchable encryption schemes are used for retrieving the original data. One of the approaches for retrieving data stored as encrypted format over cloud is keyword based search. The existing keyword based encryption schemes supports only single keyword search with order preserving encryption schemes. While considering large volume of data and users, these schemes are inefficient and cause leakage of data privacy.

To eliminate the data privacy leakage caused due to the order preserving encryption (OPE) schemes and to reduce the computational burden on the server side, a new scheme called Two Round Searchable Encryption (TRSE) with fuzzy keyword search is proposed. The proposed TRSE scheme supports multikeyword top-k retrieval over the encrypted cloud data with the help of Fuzzy keyword set. Top-k is a standard IR technique which enables fast query execution on very large indexes and makes system highly scalable. The vector space model constructed in the TRSE scheme helps to provide search accuracy and the ranking is done using fuzzy algorithm. The construction of fuzzy keyword set helps to reduce the size of vector space model thus by improves the retrieval speed from the index. Fuzzy keyword set is built based on relevance scoring.

Fuzzy keyword search greatly enhances system usability by returning the matching files when users' searching inputs exactly match the predefined keywords or the closest possible matching files based on keyword similarity semantics, when exact match fails. In this scheme, ranking is done at the user side while scoring calculation is done at the server side. Also, the similarity relevance is specific for terms and thus should be properly hidden from the cloud user. This is the main advantage of the TRSE scheme comparing to the OPE scheme which could not conceal the similarity relevance

2. RELATED WORK

Multikeyword search [2] consists of a set of strict privacy requirements for secure cloud data utilization system. It supports multikeyword query and provide result similarity ranking for effective data retrieval, instead of undifferentiated results. Privacy preserving multikeyword search technique supports two principles:

ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Biology, Engineering, Science and Technology (IJARBEST)

Vol. 1, Special Issue 5, May 2013

- Co-ordinate matching
- Inner product similarity

Co-ordinate matching includes as many matches as possible, to capture the relevance of data documents to the search query. Coordinate matching is an intermediate approach which uses the number of query keywords appearing in the document to quantify the similarity of that document to the query. Inner product similarity can be measured by the number of query keywords appearing in a document, to quantitatively evaluate such similarity measure of that document to the search query.

Boolean keyword search is associated with Searchable symmetric encryption (SSE)[3]. It allows a party to outsource the storage of its data to another party (a server) in a private manner, while maintaining the ability to selectively search over it. Boolean keyword search techniques give two solutions to the SSE scheme with respect satisfying its properties:

Sequential search without index is a practical technique for searches on encrypted data.[4]. It is desirable to store data on data storage servers such as mail servers and file servers in encrypted form to reduce security and privacy risks. But this usually implies that one has to sacrifice functionality for security. Pick a fixed-size block that is long enough to contain most words. Words that are too short or too long may be padded to a multiple of the block size with some pre-determined padding format. However, such a padding scheme would introduce space inefficiency. Secure ranked single keyword search [6] greatly enhances system usability by returning the matching files in a ranked order regarding to certain relevance criteria (e.g., keyword frequency), helps practical deployment of privacy-preserving data hosting services in Cloud Computing. This encryption scheme makes use of the advance of both crypto and IR community to design the ranked searchable symmetric encryption scheme with greater system security and usability.

3. PROPOSED METHODOLOGY

Cloud computing provides a promising pattern for data outsourcing and high-quality data services. However, concerns of sensitive information on cloud potentially cause privacy problems. Data encryption protects data security to some extent, but at the cost of compromised efficiency. Searchable symmetric encryption (SSE) allows retrieval of encrypted data over cloud. The server-side ranking based on order-preserving encryption (OPE) inevitably leaks data privacy. To eliminate the leakage of data privacy, introduce a new searchable encryption scheme called Two Round Searchable Encryption (TRSE) which retrieves files based on fuzzy keyword search. The proposed TRSE scheme with fuzzy keyword search supports multikeyword top-k retrieval over encrypted cloud data. The TRSE scheme includes vector space model for providing sufficient search accuracy. The TRSE scheme supports high security by encrypting both text files and index using different algorithms. TRSE scheme supports the relevance scoring, which includes most relevant data files.

ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Biology, Engineering, Science and Technology (IJARBEST)

Vol. 1, Special Issue 5, May 2013

Some of the multikeyword SSE schemes support only Boolean queries, ie., a file either matches or does not matches a query. The TRSE scheme takes advantage over this scenario by introducing relevance scoring. The fuzzy keyword search technique also supports relevance scoring. Based on the relevance score, files can be ranked either ascending or descending. Also TRSE scheme makes use of the advantages of both IR and cryptographic community to provide more security. In this work, the majority of computing work is done on the cloud while the user takes part in ranking. The fuzzy keyword search technique is combined with the similarity relevance scoring by using the wild card based search technique.

Consider a cloud data system consisting of data owner, data user and cloud server. Given a collection of n encrypted data files c= (F_1, F_2, ……., F_N) stored in the clod server, a predefined set of distinct keywords W= { w_1, w_2, ……, w_p} the cloud server provides the search service for the authorized user over the encrypted data c.

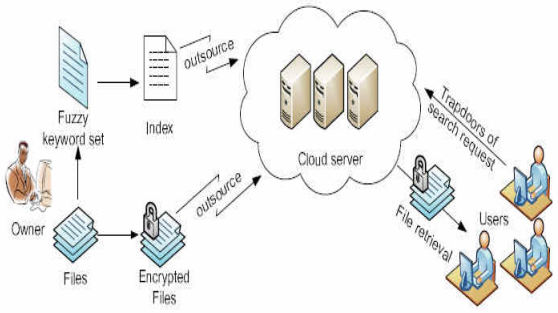


Fig 3.1 system model for the TRSE scheme with fuzzy keyword search

An authorized user types in a request to selectively retrieve data files of his/her 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: 1) if the user's searching input exactly matches the pre-set keyword, the server is expected to return the files containing the keyword; 2) if there exist typos and/or format inconsistencies in the searching input, the server will return the closest possible results based on pre-specified similarity semantics.

The proposed model consists of Initialization and Retrieval. The Initialization phase includes Setup and IndexBuild. The Setup stage involves the secure initialization, while the IndexBuild stage involves operations on plaintext. For security concerns, the vast majority of work should only be done by the data owner. The Index build phase includes similarity relevance technique with fuzzy keyword set. The Retrieval phase involves TrapdoorGen, ScoreCalculate, and Rank, in which the data user and the cloud server are involved. As a result of the limited computing power on the user side, the computing work should be left to server side as much as possible. Meanwhile, the confidentiality privacy of sensitive information cannot be violated.

In Intialization phase The data owner calls KeyGen(λ) to generate secret key SK and public key set PK for the homomorphic encryption scheme. Then the data owner assigns SK to the authorized data users. The data owner extracts the collection of l keywords, W={ w_1, w_2, …….,w_l },And their TF and IDF values out of the collection of n files, C={

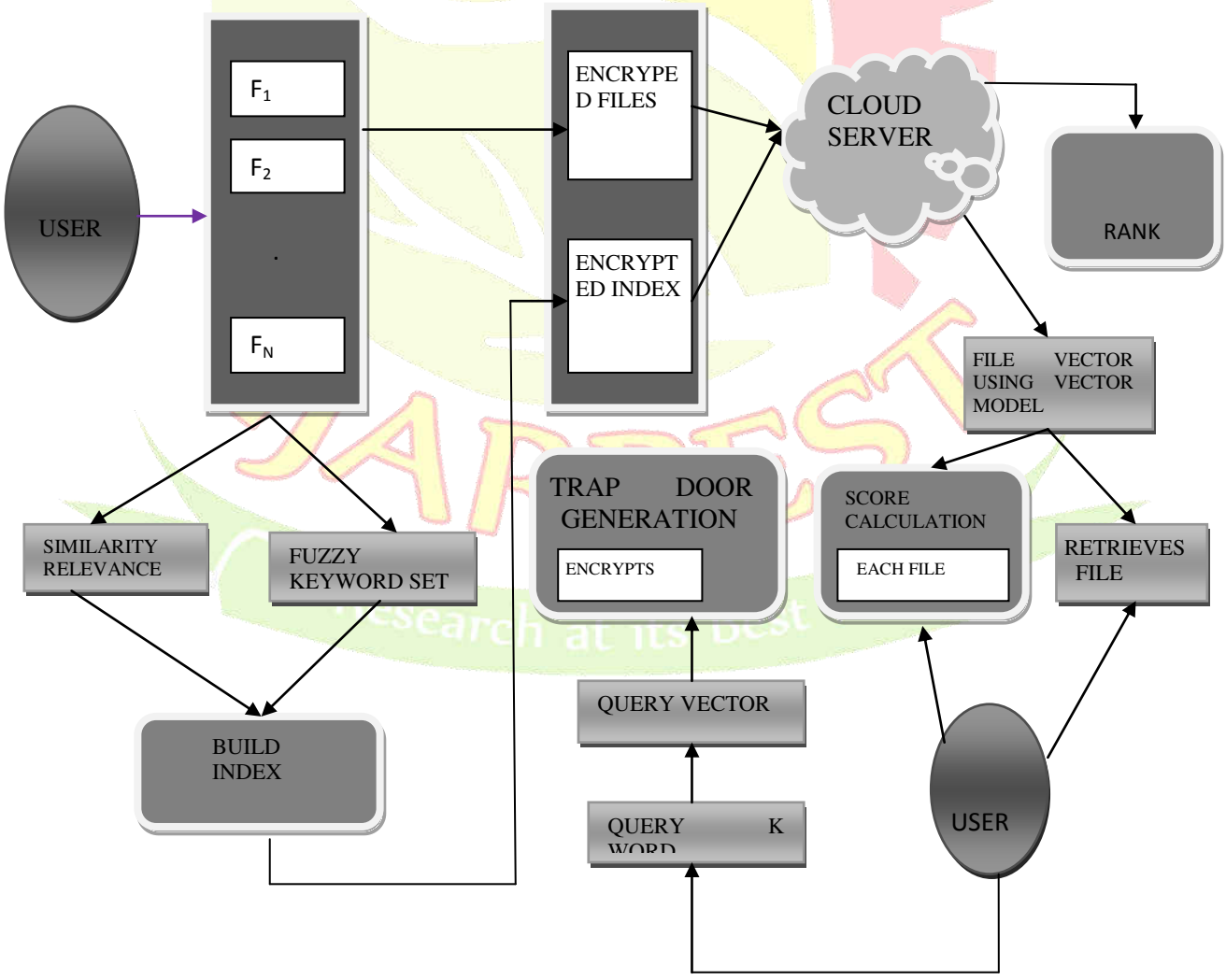
ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Biology, Engineering, Science and Technology (IJARBEST)

Vol. 1, Special Issue 5, May 2013

$f_1, f_2, \ldots\ldots, f_n$ }. For each file $f_i \epsilon C$, the data owner builds a (l+1)- dimensional vector $v_{i=}$ {$id_i, t_{i,1}, t_{i,2}, \ldots.t_{i,l}$ },where $t_{i,j}$ = tf-$idf_{wj,fi}$ ($1 \leq j \leq l$). The searchable index $I=\{v_i | 1 \leq i \leq n\}$. The data owner encrypts the searchable index I to secure searchable index I.The data owner encrypts C= {$f1, f2, \ldots\ldots, f_n$ } into C' with other cryptographic schemes and then outsource C' and I' to the cloud server.

In Retrieval phase The data user generates a set of keywords REQ= {w_1', w2'… w_s' } to search , and then the query vector T_w= {$m_1, m_2, \ldots.., m_l$} is generated in which m_i= $1(1 \leq i \leq l)$. T_w is encrypted and then sends to the cloud server. For each file vector v_j' $(0 \leq j \leq n)$ in I',the cloud server computes the inner product P_j'= v_j'[1:l]. The secure trapdoor with modular reduction and then compresses and returns the result vector N' to the data user. The data user decrypts N' into N. Then TOPKSELECT(N,k) is invoked to get the top-k highest scoring files identifiers and sends it to the cloud server. he cloud server returns the encrypted k files to the data user.



ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Biology, Engineering, Science and Technology (IJARBEST)

Vol. 1, Special Issue 5, May 2013

Fig 3.2 PROBLEM WORKFLOW

3.1 FRAMEWORK OF TRSE WITH FUZZY KEYWORD SEARCH

A type of search that will find matches even when user misspells words.

SETUP-The data owner generates the secret key and the public keys using security parameter λ for the homomorphic encryption scheme.

INDEXBUILD- The data owner builds the secure searchable index from the file collection C. To form search index I from C, Information Retrieval community (IR) techniques are used to build index by using similarity relevance technique and fuzzy keyword set is build using edit distance technique in wild-card based search. The final index is built by combining these two techniques. Finally I is encrypted to I' using PK.

TRAPDOORGEN- The data user generates secure trapdoor from his request REQ. Vector Tw is built from the user's multi keyword request. Encrypt using public key PK Built the secure trapdoor Tw'.

SCORECALCULATE- Cloud server receives the secure trapdoor Tw', Computes the scores of each file I' with Tw' Returns the encrypted result vector N back to the data user.

RANK-The data user decrypts the vector N using SK. Requests and gets the files with top-k scores.

3.2 FUZZY KEYWORD SEARCH

Here the fuzzy keyword set is built using edit distance with wild-card based technique.

3.2.1 EDIT DISTANCE

Edit distance computes the distance between two keywords k1, k2 by using ed(k1,k2). Edit distance is used to measure the keyword similarity. If any dissimilar occurs, three operations can be carried out:

a) Substitution: substitute's one character with the other character. b) Insertion: inserting one character into other character. c) Deletion: deleting one character from the other character. It defines the fuzzy keyword search as follows: Given a collection of data files c= (df1, df2, …….. , dfn) are stored in cloud and set of keywords (k1,k2,……. ,kn). When the user sends request this technique searches with edit distance and keyword.

3.2.2 WILD-CARDBASED DISTANCE

In wild-card based technique, all the variants of keywords to be listed when an operation is performed at the same position. Based on the above approach, we use a wild card to denote the edit operations performed at the same position. This technique edits distance to solve the problems. It includes the following steps:

It builds an index with each keyword k. To build index data owner computes f(sk , k). Construct the secret key sk. This sk is shared among the data owner and

ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Biology, Engineering, Science and Technology (IJARBEST)

Vol. 1, Special Issue 5, May 2013

user if he is an authorized user. Searching can be done with secret key sk, keyword k. Compare the secret key sent by the user and existed key at the data owner. If both are same, returns the requested file. For example, for the keyword FEVER with the pre-set edit distance 1, its wildcard based fuzzy keyword set can be constructed as FEVER,1 = {FEVER, *FEVER, *EVER, FE*VER, F*VER, FEV*ER, FEV*R, FEVE*R, FEVE*, FEVER*}.

3.3 HOMOMORPHIC ENCRYPTION

Homomorphic encryption is a form of encryption which allows specific types of computations to be carried out on cipher text and obtain an encrypted result which decrypted matches the result of operations performed on the plaintext. For instance, one person could add two encrypted numbers and then another person could decrypt the result, without either of them being able to find the value of the individual numbers. This is a desirable feature in modern system architectures.

4. PERFORMANCE ANALYSIS

The performance of our scheme is evaluated regarding the time cost of fuzzy set construction, the time and storage cost of index construction, the search time of the listing approach and the wild-card based approach.

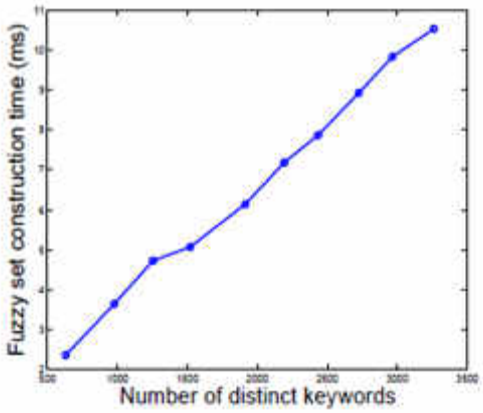


Fig 4.1 Edit distance d=1

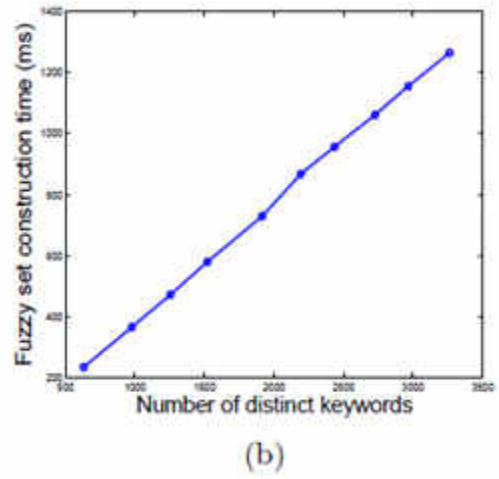


Fig 4.1 Edit distance d=2

The fuzzy set construction time using wild-card based approach a) edit distance d=1 b) edit distance d=2

5. CONCLUSIONS AND FUTURE WORK

ISSN (ONLINE): 2395-695X
ISSN (PRINT): 2395-695X

International Journal of Advanced Research in Biology, Engineering, Science and Technology (IJARBEST)

Vol. 1, Special Issue 5, May 2013

The two round searchable encryption scheme supports multi keyword retrieval with fuzzy keyword search. This search is more efficient and secure than other searchable encryption schemes. Results show that the computation overhead on the client side is reduced. It also performs well for calculating ranks by using relevant scores. The TRSE scheme employs the fully homomorphic encryption which fulfills the security requirements of multi keyword top-K retrieval over the encrypted cloud data and the fuzzy keyword set is developed based on similarity relevance that improves the retrieving speed of the document from cloud. Security analysis shows that access pattern and search patterns are secured because of the encryption of different keywords in same queries and are independent. As future work the TRSE scheme using Fuzzy keyword search with symbol based trie traverse technique can be used to improve the retrieving speed of files stored in cloud server.

REFERENCES

[1] Boneh, G. Crescenzo, R. Ostrovsky, and G. Persiano, "Public- Key Encryption with Keyword Search," Proc. Int'l Conf. Theory and Applications of Cryptographic Techniques (Eurocrypt), 2004.

[2] Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-Preservingmulti-keyword Ranked Search over Encrypted Cloud Data," *Proc. IEEE INFOCOM, 2011.*

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

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

[5] Swaminathan, Y. Mao, G.-M. Su, H. Gou, A.L. Varna, S. He, M. Wu, and D.W. Oard," Confidentiality-Preserving Rank-Ordered Search," Proc. Workshop Storage Security and Survivability, 2007.

[6] Wang, N. Cao, J. Li, K. Ren, and W. Lou, "Secure Ranked Keyword Search over Encrypted Cloud Data," Proc. IEEE 30th Int'l Conf. Distributed Computing Systems (ICDCS), 2010.

[7] Wang , Coll. of Comput., Nat. Univ. of Defense Technol.,Changsha, China Shaojing Fu,Ming Xu, " Privacy preserving fuzzy keyword search over encrypted cloud data", Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on (Volume:1)