A SECURE IMPLEMENTATION OF MULTI KEYWORD SEARCH IN ENCRYPTED CLOUD DATA WITH RANKING

G.Archana[1], S.Tharun Reddy[2]
[1]M.Tech Student, [2]Assistant Professor, Department of CSE, Vaagdevi College of Engineering,
Bollikunta, Warangal, Telangana, India

Abstract:* As cloud computing become more flexible & effective in terms of economy, data owners are motivated to outsource their complex data systems from local sites to commercial public cloud. But for security of data, sensitive data has to be encrypted before outsourcing, which overcomes method of traditional data utilization based on plaintext keyword search. Considering the large number of data users and documents in cloud, it is necessary for the search service to allow multi-keyword query and provide result similarity ranking to meet the effective data retrieval need. Retrieving of all the files having queried keyword will not be affordable in pay as per use cloud paradigm. In this paper, we propose the problem of Secured Multikeyword search (SMS) over encrypted cloud data (ECD), and construct a group of privacy policies for such a secure cloud data utilization system. From number of multi-keyword semantics, we select the highly efficient rule of coordinate matching, i.e., as many matches as possible, to identify the similarity between search query and data , and for further matching we use inner data correspondence to quantitatively formalize such principle for similarity measurement. We first propose a basic Secured multi keyword ranked search scheme using secure inner product computation, and then improve it to meet different privacy requirements. The Ranked result provides top k retrieval results. Also we propose an alert system which will generate alerts when un-authorized user tries to access the data from cloud, the alert will generate in the form of mail and message.

***Keywords:* Encryption, Inner product similarity, Multi-keyword search, ranking.**

I. INTRODUCTION

The cloud makes it possible for you to access your information from anywhere at any time. While a traditional computer setup requires you to be in the same location as your data storage device, the cloud takes away that step. The cloud removes the need for you to be in the same physical location as the hardware that stores your data. Your cloud provider can both own and house the hardware and software necessary to run your home or business applications. This is especially helpful for businesses that cannot afford the same amount of hardware and storage space as a bigger company. Small companies can store their information in the cloud, removing the cost of purchasing and storing memory devices. Additionally, because you only need to buy the amount of storage space you will use, a business can purchase more space or reduce their subscription as their business grows or as they find they need less storage space. Each provider serves a specific function, giving users more or less control over their cloud depending on the type. When you choose a provider, compare your needs to the cloud services available. The information housed on the cloud is often seen as valuable to individual with malicious intent. There is a lot of personal information and potentially secure data that people store on their computers, and this information is now being transferred to the cloud. This makes it critical for you to understand the security measures that your cloud provider has in place, and it is equally important to take personal precautions to secure your data. The multi-keyword retrieval over encrypted cloud data achieves high security and privacy.

II. RELATED WORK

As Cloud Computing becomes prevalent, sensitive information are being increasingly centralized into the cloud. For the protection of data privacy, sensitive data has to be encrypted before outsourcing, which makes effective data utilization a very challenging task. Ranked search greatly enhances system usability by returning the matching files in a ranked order regarding to certain relevance criteria (e.g., keyword frequency), thus making one step closer towards practical deployment of privacy-preserving data hosting services in Cloud Computing. We first give a straightforward yet ideal construction of ranked keyword search under the state-of-the-art searchable symmetric encryption (SSE)[1].By determining the most common English words and phrases since the beginning of the sixteenth century, we obtain a unique large-scale view of the evolution of written text. We find that the most common words and phrases in any given year had a much shorter popularity lifespan in the sixteenth century than they had in the twentieth century. The task of compiling a monograph on corpus linguistics must be faced up with the problem that there is at present no consensus among linguists. It is generally admitted that the mere fact of quoting authentic examples of language use is not a sufficient condition for a piece of research to qualify' as corpus linguistic. they provide provable secrecy for encryption, in the sense that the untrusted server cannot learn anything about the plaintext when only given the cipher text; they provide query isolation for searches, meaning that the untrusted server cannot learn anything more about the plaintext than the search result; they provide controlled searching, so that the untrusted server cannot search for an arbitrary word without the user's authorization; they also support hidden queries, so that the user may askthe untrusted server to search for a secret word without revealing the word to the server. Privacy preserving multi-keyword ranked search over encrypted cloud data (MRSE). We establish a set of strict privacy requirements for such a secure cloud data utilization system. Among various multi-keyword semantics, we choose the efficient similarity measure of “coordinate matching”, i.e., as many matches as possible, to capture the relevance of data documents to the search query. We further use “inner product similarity” to quantitatively evaluate such similarity measure. Fuzzy keyword searchgreatly 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 our solution, we exploited it distance to quantify keywords similarity and develop two advanced techniques on constructing fuzzy keyword sets, which achieve optimized storage and representation overheads. A new symbol-based tire-traverse searching scheme, where a multi-way tree structure is built up using symbols transformed from the resulted fuzzy keyword sets is constructed. If the user is actually interested in documents containing each of several keywords the user must either give the server capabilities for each of the keywords individually or rely on an intersection calculation to determine the correct set of documents, or alternatively, the user may store additional information on the server to facilitate such searches. Neither solution is desirable; the former enables the server to learn which documents match each individual keyword of the conjunctive search and the latter results in exponential storage if the user allows for searches on every set of keywords.

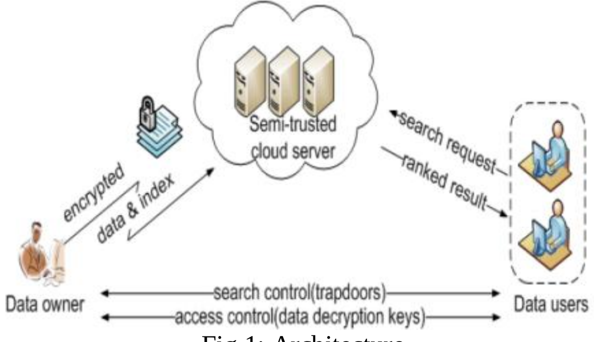

Semi-trusted cloud server
search request
ranked result
encrypted data & index
Data owner
Data users
search control(trapdoors)
access control(data decryption keys)

Fig 1: Architecture

III. FRAMEWORK AND PRIVACY REQUIREMENTS FOR MRSE

In this section, we define the framework of multi-keyword ranked search over encrypted cloud data (MRSE) and establish various strict system-wise privacy requirements for such a secure cloud data utilization system.

A. MRSE Framework: For easy presentation, operations on the data documents are not shown in the framework since the data owner could easily employ the traditional symmetric key cryptography to encrypt and then outsource data. With focus on the index and query, the MRSE system consists of four algorithms as follows.

• Setup(1ℓ) Taking a security parameter ℓ as input, the data owner outputs a symmetric key as SK.

• BuildIndex(F, SK) Based on the dataset F, the data owner builds a searchable index I which is encrypted by the symmetric key SK and then outsourced to the cloud server. After the index construction, the document collection can be independently encrypted and outsourced.

• Trapdoor(Wf) With t keywords of interest in Wf as input, this algorithm generates a corresponding trapdoor TWf.

• Query(TWf, k, I) When the cloud server receives a query request as (TWf, k), it performs the ranked search on the index I with the help of trapdoor TWf, and finally returns FWf, the ranked id list of top-k documents sorted by their similarity with Wf. Neither the search control nor the access control is within the scope of this paper. While the former is to regulate how authorized users acquire trapdoors, the later is to manage users’ access to outsourced documents.

B. Privacy Requirements for MRSE: The representative privacy guarantee in the related literature, such as searchable encryption, is that the server should learn nothing but search results. With this general privacy description, we explore and establish a set of strict privacy requirements specifically for the MRSE framework. As for the data privacy, the data owner can resort to the traditional symmetric key cryptography to encrypt the data before outsourcing, and successfully prevent the cloud server from prying into the outsourced data. With respect to the index privacy, if the cloud server deduces any association between keywords and encrypted documents from index, it may learn the major subject of a document, even the content of a short document. Therefore, the searchable index should be constructed to prevent the cloud server from performing such kind of association attack. While data and index privacy guarantees are demanded by default in the related literature, various search privacy requirements involved in the query procedure are more complex and difficult to tackle as follows.

Keyword Privacy As users usually prefer to keep their search from being exposed to others like the cloud server, the most important concern is to hide what they are searching, i.e., the keywords indicated by the corresponding trapdoor. Although the trapdoor can be generated in a cryptographic way to protect the query keywords, the cloud server could do some statistical analysis over the search result to make an estimate. As a kind of statistical information, document frequency (i.e., the number of documents containing the keyword) is sufficient to identify the keyword with high probability. When the cloud server knows some background information of the dataset, this keyword specific information may be utilized to reverse-engineer the keyword. Trapdoor Unlink abilityThe trapdoor generation function should be a randomized one instead of being deterministic. In particular, the cloud server should not be able to deduce the relationship of any given trapdoors, e.g., to determine whether the two trapdoors are formed by the same search request. Otherwise, the deterministic trapdoor generation would give the cloud server advantage to accumulate frequencies of different search requests regarding different keyword(s), which may further violate the aforementioned keyword privacy requirement. So the fundamental protection for trapdoor unlink ability is to introduce sufficient non-determinacy into the trapdoor generation procedure.Access Pattern Within the ranked search, the access pattern is the sequence of search results where every search result is a set of documents with rank order. Specifically, the search result for the query keyword set Wf is denoted as FWf, consisting of the id list of all documents ranked by their relevance to Wf. Then the access pattern is denoted as (FWf1 , FWf2 , . . .) which are the results of sequential searches. Although a few searchable encryption works, e.g., has been proposed to utilize private information retrieval (PIR) technique, to hide the access pattern, our proposed schemes are not designed to protect the access pattern for the efficiency concerns. This is because any PIR based technique must "touch" the whole dataset outsourced on the server which is inefficient in the large scale cloud system.

IV. SECURITY ANALYSIS

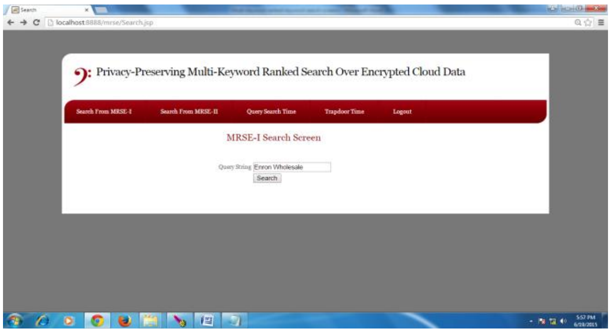
We analyze our proposed scheme according to the predefined privacy requirements in design goals:

1) Index Confidentiality. In our proposed scheme, i I and w T are obfuscated vectors, which means the cloud servercan not infer the original data vector and the query vector without the secret key SK. The cloud server cannot deduce TF values from the result relevance scores. In other word, the index confidentiality is protected.

2) Trapdoor Unlinkability. The trapdoor of query vector is generated from a random splitting operation, which means the same search requests are transformed into different query trapdoors. And thus, the query unlinkability is well preserved.

3) Keyword Privacy. In the known background scheme, the cloud server is supposed to have more knowledge, 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 is designed to obscure the TF distributions of keywords with the dummy values. That is to say, the keyword privacy is protected.

V. IMPLEMENTATION RESULTS

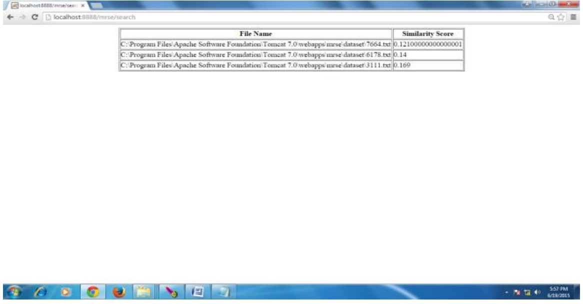


Privacy-Preserving Multi-Keyword Ranked Search Over Encrypted Cloud Data

MRSE-I Search Screen

Query String: Enron Wholesale

Fig 2: Searching data via keywords



File Name	Similarity Score
C:\Program Files\Apache Software Foundation\Tomcat 7.0\webapps\mrse\dataset\7664.txt	0.12100000000000001
C:\Program Files\Apache Software Foundation\Tomcat 7.0\webapps\mrse\dataset\6178.txt	0.14
C:\Program Files\Apache Software Foundation\Tomcat 7.0\webapps\mrse\dataset\3111.txt	0.169

Fig 3: Search Results

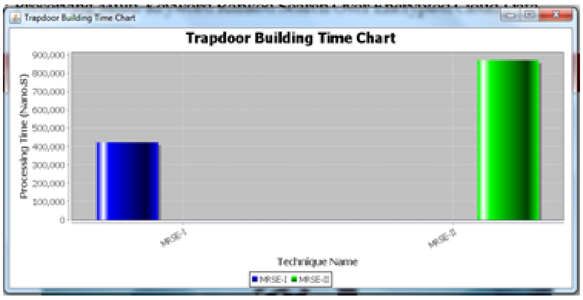


Fig 4: Trapdoor Building time chart

VI. CONCLUSION

In this paper, for the first time we define and solve the problem of multi-keyword ranked search over encrypted cloud data, and establish a variety of privacy requirements. Among various multi-keyword semantics, we choose the efficient similarity measure of "coordinate matching", i.e., as many matches as possible, to effectively capture the relevance of outsourced documents to the query keywords, and use "inner product similarity" to quantitatively evaluate such similarity measure. For meeting the challenge of supporting multi-keyword semantic without privacy breaches, we propose a basic idea of MRSE using secure inner product computation. Then we give two improved MRSE schemes to achieve various stringent privacy requirements in two different threat models. Thorough analysis investigating privacy and efficiency guarantees of proposed schemes is given, and experiments on the real-world dataset show our proposed schemes introduce low overhead on both computation and communication. In our future work, we will explore supporting other multi-keyword semantics (e.g., weighted query) over encrypted data and checking the integrity of the rank order in the search result

REFERENCES

[1] D. Boneh, E. Kushilevitz, R. Ostrovsky, and W. E. S. III, "Public key encryption that allows pir queries," in Proc. of CRYPTO, 2007.

[2] P. Golle, J. Staddon, and B. Waters, "Secure conjunctive keyword search over encrypted data," in Proc. of ACNS, 2004, pp. 31–45.

[3] L. Ballard, S. Kamara, and F. Monrose, "Achieving efficient conjunctive keyword searches over encrypted data," in Proc. of ICICS, 2005.

[4] D. Boneh and B. Waters, "Conjunctive, subset, and range queries on encrypted data," in Proc. of TCC, 2007, pp. 535–554.

[5] R. Brinkman, "Searching in encrypted data," in University of Twente, PhD thesis, 2007.

[6] Y. Hwang and P. Lee, "Public key encryption with conjunctive keyword search and its extension to a multi-user system," in Pairing, 2007.

[7] J. Katz, A. Sahai, and B. Waters, "Predicate encryption supporting disjunctions, polynomial equations, and inner products," in Proc. of EUROCRYPT, 2008

[8] I. H. Witten, A. Moffat, and T. C. Bell, Managing gigabytes: Compressing and indexing documents

and images, Morgan Kaufmann Publishing, San Francisco, May 1999.

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

[10] D. Song, D. Wagner, and A. Perrig, ―Practical techniques for searches on encrypted data,‖ in Proc. of IEEE Symposium on Security and Privacy'00, 2000.

[11] E.-J. Goh, ―Secure indexes,‖ Cryptology ePrint Archive, Report 2003/216, 2003, http://eprint.iacr.org/.

G.ARCHANA Currently doing M.Tech in Computer Science & Engineering at Vaagdevi College of Engineering, Bollikunta, Warangal, India and her Research area includes Data Mining ,Cloud Computing, Network Security etc.,

S.Tharun Reddy is 5+ years experienced Assistant Professor in the Department of Computer Science & Engineering, Vaagdevi College of Engineering, Bollikunta, Warangal, India and his Research area includes Data Mining ,Cloud Computing, Network Security, Design & Analysis of Algorithm, Stenography etc.,