www.ioirp.com
International Journal of Innovative Research in Computer Science and Engineering (IJIRCSE)
ISSN: 2394-6364, Volume – 1, Issue – 6. May 2015

SEMANTIC MULTI-KEYWORD SEARCH OVER ENCRYPTED CLOUD DATA

KrishnapriyaKailas
M.E.(C.S.E), Maharaja Prithvi Engineering College, Avinashi
krishna8969@gmail.com

Abstract— With the growing popularity of cloud computing, data owners are motivated to outsource the complex data management systems from local sites to commercial public cloud for reduced management cost and the ease of access. To ensure the safety of stored data, it has to be encrypted before outsourcing. In this scheme, the multi-keyword ranked search over encrypted cloud data is defined and solved. This search uses the feature of coordinate matching, i.e., as many matches as possible, to capture the relevance of data documents to the search query and inner product similarity to quantitatively evaluate such similarity measure. Two secure schemes are proposed to meet privacy requirements in two threat models of known cipher text model and known background model. The existing solutions depended entirely on the submitted query keyword and didn't consider the semantics of keyword. Thus the search schemes are not intelligent and also omit some semantically related documents. In view of the deficiency, as an attempt, a semantic multi-keyword ranked search scheme over the encrypted cloud data, which simultaneously meets a set of strict privacy requirements, is proposed. It uses the Latent Semantic Analysis to reveal relationship between terms and documents. The relationship between terms is automatically captured. The solution could return not only the exactly matched files but also the files including the terms semantically related to the query keyword. Also this scheme can upload any type of files including video files in encrypted cloud. Experimental evaluation demonstrates the efficiency and effectives of the scheme.

Keywords— *Latent Semantic search; Multi-keyword; Rank; Secure; Cloud data*

I. INTRODUCTION

Cloud Computing enables cloud customers to enjoy the on-demand high quality applications and services from a centralized pool of configurable computing resources. This new computing model can relieve the burden of storage management, allow universal data access with independent geographical locations, and avoid capital expenditure on hardware, software, and personnel maintenances, etc. As cloud computing becomes mature, lots of sensitive data is considered to be centralized into the cloud servers, e.g. personal health records, secret enterprise data, government documents, etc. The straightforward solution to protect data privacy is to encrypt sensitive data before being outsourced. Unfortunately, data encryption, if not done appropriately, may reduce the effectiveness of data utilization. Typically, a user retrieves files of interest to him/her via keyword search instead of retrieving back all the files. Such keyword-based search technique has been widely used in daily life, e.g. the Google plaintext keyword search. However, technologies are invalid after the keywords are encrypted.

In recent years, searchable encryption (SE) techniques have been developed for secure outsourced data search. In the existing system, the problem of multi-keyword ranked search over encrypted cloud data (MRSE) is solved and establishes a set of strict privacy requirements for such a cloud data utilization system. The search uses the efficient principle of "coordinate matching", i.e., as many matches as possible, to capture the similarity between search query and data documents, and further use "inner product similarity" to quantitatively formalize such principle for similarity measurement. We first propose a basic MRSE scheme using secure inner product computation, and then significantly improve it to meet different privacy requirements in two levels of threat models. The two threat models with different attack capabilities are Known cipher model and Known background model. Thorough analysis investigating privacy and efficiency guarantees of proposed schemes is given, and experiments on the real-world dataset further show proposed schemes indeed introduce low overhead on computation and communication.

The existing solutions depended entirely on the submitted query keyword and didn't consider the semantics of keyword. That means they don't support searching for different variants of the query word, which is a significant drawback and greatly affects data usability and user experience. Thus the search schemes are not intelligent and also omit some semantically related documents. Also the existing solutions do not support all types of file formats for uploading in encrypted cloud environment.

In view of the deficiency, as an attempt, this paper propose a semantic multi-keyword ranked search scheme over the encrypted cloud data, which simultaneously meets a set of strict privacy requirements. It uses a "Latent Semantic Analysis" to reveal relationship between terms and documents. The latent semantic analysis takes advantage of implicit higher-order structure in the association of terms with documents and adopts reduced-dimension vector space to represent words and documents. Thus,

the relationship between terms is automatically captured. The solution could return not only the exactly matched files, but also the files including the terms semantically related to the query keyword. Also, the proposed scheme can upload any type of files including video files in the encrypted cloud. But, in the case of video files, exact encryption cannot be done instead, it can be stored in a hidden location and can access when needed. Experimental evaluation demonstrates the efficiency and effectives of the scheme.

II. PROBLEM FORMULATION

A. *System Model*

The system model can be considered as three entities, as depicted in Fig .1: the data owner, the data user and the cloud server. Data owner has a collection of data documents D = { , ,..., } dd d 1 2 m . A set of distinct keywords W== { , ,..., } ww w 1 2 n is extracted from the data collection D . The data owner will firstly construct an encrypted searchable index I from the data collection D .Then, the data owner upload both the encrypted index I and the encrypted data collection C to the cloud server. Data user provides t keywords for the cloud server. The cloud server only sends back top-l files that are most relevant to the search query.

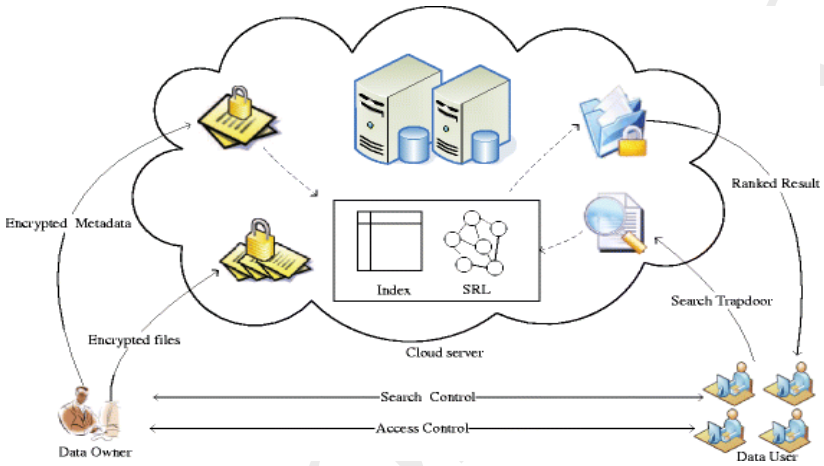

Fig.1. Architecture

B. *Threat Model*

Cloud server is considered as "honest-but-curious" in our model, which is consistent with the most related works on searchable encryption. Specifically, cloud server acts in an "honest" fashion and correctly follows the designated protocol specification. However, it is "curious" to infer and analyze data (including index) in its storage and message flows received during the protocol so as to learn additional information. Based on what information cloud server knows, we consider two levels of threat models as follows.

- *Known Ciphertext Model:* In this model, cloud server is supposed to only know encrypted dataset C and searchable index I, both of which are outsourced from data owner.

- *Known Background Model:* In this stronger model, cloud server is supposed to possess some backgrounds on the dataset, such as the subject and its related statistical information, in addition to what can be accessed in known ciphertext model. As an instance of possible attacks in this case, cloud server could utilize document frequency or keyword frequency to identify keywords in the query example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

C. *Design Goals*

To activate the ranked search for effective utilization of outsourced cloud data under the aforementioned model, the system should be designed by considering the security considerations also. The system is expected to give the following security and performance guarantees as follows:

- *Latent Semantic Search:* The statistical techniques to estimate the latent semantic structure is used and get rid of the obscuring "noise"

- *Multi-keyword Ranked Search:* It supports both multi-keyword query and support result ranking. This includes searching all types of files including video files.
- *Privacy-Preserving:* Our scheme is designed to meet the privacy requirement and prevent the cloud server from learning additional information from index and trapdoor.

1. Index Confidentiality. The TF values of keywords are stored in the index. Thus, the index stored in the cloud server needs to be encrypted;
2. Trapdoor Unlinkability. The cloud server should not be able to deduce relationship between trapdoors.
3. Keyword Privacy. The cloud server could not discern the keyword in query, index by analyzing the statistical information like term frequency.

D. Notations and Preliminaries

D --the plaintext document collection, denoted as a set of n dataλ documents D = { , ,..., } dd d 1 2 m

C --the encrypted document collection stored in the cloud server, denotedλ as C= { , ,..., } cc c 1 2 m .

W=--the dictionary, the keyword set composing of m keyword, denotedλ as W== { , ,..., } ww w 1 2 n .

I --the searchable index associatedC , denoted as 1 2 (, ,...,) m II I .λ i j ,

tf --the term frequency, the i-th term appears times in the j-th document.λ

A′[j] --the data vector for document d j where the element A′[i][j] represents theλ term frequency i j , tf of the corresponding keyword W=i in document d j .

Q --the query vector indicating the keywords of interest where eachλ bit Q[ ] {0,1} j ∈ represents the existence of the corresponding keyword in the .%query W

Latent Semantic Analysis: In information retrieval, latent semantic analysis is a solution for discovering the latent semantic relationship. It adopts singular-value decomposition, which is abbreviated as SVD to find the semantic structure between terms and documents. In this paper, the term-document matrix consists of n rows, each of which represents the data vector for each file,

$$A' = (A'[1] \dots A'[j] \dots A'[m]) \quad (1)$$

as depicted in the Eq.1.Then, we take a large term-document matrix and decompose it into a set of k , orthogonal factors from which the original matrix can be approximated by linear combination . For example, a term-document matrix named A′ can be decomposed into the product of three other matrices:

$$A' = U' * V' * S' \quad (2)$$

such that U′ and V′ have orthonormal columns, S′ is diagonal. We choose previous k columns of S′ , and then deleting the corresponding columns of U′ and V′ respectively.The result is a reduced model:

$$A = U' . V' . S' \sim A' \quad (3)$$

Secure k-NN: In order to compute the inner product in a privacy-preserving method, we will adapt the secure k -nearest neighbor scheme. This splitting technique is secure against known-plaintext attack, which is roughly equal in security to a d-bit symmetric key.

III. PROPOSED SCHEME

According to the above definition about LSA, the data owner builds a term-document matrix A′ . We reduce the dimensions of the original matrix A′ to get a new matrix A which is calculated the best "reduced-dimension" approximation to the original term-document matrix. Specially, A[j] denotes the j -th column of the matrix A .

Setup: The data owner generates a n + 2 -bit vector as X and two (n+ 2) (n+2) invertible matrices { M1 ,M2} .The secret key SK is the form of a 3-tuple as { X,M1 ,M2}

*BuildIndex(A', SK):*The data owner extracts a term-document matrix A′ . Following, we multiply these three matrices to get the result matrix A .Taking privacy into consideration, it is necessary that the matrix A is encrypted before outsourcing. After

applying dimension-extending, the original A[j] is extended to (n + 2) -dimensions, instead of n . Namely, the (n +1) -th entry in A[j] is set to a random number j ε , and the (n + 2)-th entry in A[j] is set to 1 during the dimension extending.

Trapdoor(W)vector Q is generated. The (n +1) -th entry in Q is set to a random number 1, and then scaled by a random number $r \neq 0$, and the (n + 2) -th entry in Q is set to is%another random number t.

IV. PERFORMANE ANALYSIS

F-measure that combines precision and recall is the harmonic mean of precision and recall[8]. Here, we adopt F-measure to weigh the result of our experiments. 2 precision recall F = precision+ recall · · (5) For a clear comparison, our proposed scheme attains 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. .

V. CONCLUSION

The multi-keyword ranked search scheme over encrypted cloud data is solved as an existing approach. This search uses the feature of "coordinate matching", i.e., as many matches as possible, to capture the relevance of data documents to the search query and "inner product similarity" to quantitatively evaluate such similarity measure. Two secure schemes are proposed to meet privacy requirements in two threat models of known cipher text model and known background model. The existing solutions depended entirely on the submitted query keyword and didn't consider the semantics of keyword. In view of the deficiency, as an attempt, this paper propose a semantic multi-keyword ranked search scheme over the encrypted cloud data, which simultaneously meets a set of strict privacy requirements. It uses the "Latent Semantic Analysis" to reveal relationship between terms and documents. The relationship between terms is automatically captured. The solution could return not only the exactly matched files, but also the files including the terms semantically related to the query keyword. Also, the proposed scheme can upload any type of files including video files in the encrypted cloud. Experimental evaluation demonstrates the efficiency and effectives of the scheme.

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, Jan, 2014.

[2] Buyrukbilen, S. Grad. Center, City Univ. of New York, Bakiras, S. " Privacy- Preserving Ranked Search on Public-Key Encrypted Data",2013 IEEE International Conference on High Performance Computing and Communications

[3] C. Wang, N. Cao, K. Ren, and W. Lou, "Enabling Secure and Efficient Ranked Keyword Search over Outsourced Cloud Data," IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 8, pp. 1467-1479, Aug. 2012.

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

[5] E. Shen, E. Shi, and B. Waters, "Predicate Privacy in Encryption Systems," Proc. Sixth Theory of Cryptography Conf. Theory of Cryptography (TCC), 2009.

[6] F. Bao, R. H. Deng, X. Ding, and Y. Yang. Private query on encrypted data in multi-user settings. In ISPEC'08, pages 71–85, Berlin, Heidelberg, 2008. Springer-Verlag.

[7] Jiadi Yu," Toward Secure Multi-keyword Top-k Retrieval over Encrypted Cloud Data" IEEE transactions on dependable and secure computing,vol. 10, no. 4, July/August,2013.