International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

IJIEMR Transactions, online available on 16th July 2017. Link :
http://www.ijiemr.org/downloads.php?vol=Volume-6&issue=ISSUE-5

Title: Search Engine Based Ranking Of Polymorphic K-Search On Ciphered Data.

Volume 06, Issue 05, Page No: 1877 – 1882.

Paper Authors

***P.DIVYA, G.VENKATESH.**

* Dept of CSE, Shri Shiridi Sai Institute of Science and Engineering.

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

SEARCH ENGINE BASED RANKING OF POLYMORPHIC K-SEARCH ON CIPHERED DATA

***P.DIVYA,**G.VENKATESH**

* PG Scholar, Dept of CSE, Sri Venkateshwara Engineering College, Suryapet, T.S, India.

**Assistant Professor, Dept of CSE, Sri Venkateshwara Engineering College, Suryapet, T.S, India.

ABSTRACT:

As so much advantage of cloud computing, more and more data owners 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 semantic multi-keyword ranked search scheme over the encrypted cloud data, which simultaneously meets a set of strict privacy requirements. Firstly, we utilize the "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 ("semantic structure") and adopts a reduced-dimension vector space to represent words and documents. Thus, the relationship between terms is automatically captured. Secondly, our scheme employ secure "k-nearest neighbor (k-NN)" to achieve secure search functionality. The proposed scheme could return not only the exact matching files, but also the files including the terms latent semantically associated to the query keyword. Finally, the experimental result demonstrates that our method is better than the original MRSE scheme.

Keywords: sensitive data, multi-keyword ranked search, latent semantic

INTRODUCTION:

Due to the rapid expansion of data, the data owners tend to store their data into the cloud to release the burden of data storage and maintenance [1]. However, as the cloud customers and the cloud server are not in the same trusted domain, our outsourced data may be under the exposure to the risk. Thus, before sent to the cloud, the sensitive data needs to be encrypted to protect for data privacy and combat unsolicited accesses. Unfortunately, the traditional plaintext search methods cannot be directly applied to the encrypted cloud data any more. The traditional information retrieval (IR) has already provided multi-keyword ranked search for the data user. In the same way, the cloud server needs provide the data user with the similar function, while protecting data and search privacy. It is meaningful storing it into the cloud server only when data can be easily searched and utilized. In the literature, searchable encryption techniques are able to provide secure search over encrypted data for users. They build a searchable inverted index that stores a list of mapping from keywords to the corresponding set of files which contain this keyword. When data users input a keyword, a trapdoor is generated for this keyword and then

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

submitted to the cloud server. Upon receiving the trapdoor, the cloud server executes comparison between the trapdoor and index, and finally returns the data users all files that contain this keyword. 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., 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×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., 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., 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 this scheme.In this paper, we will solve the problem of multi-keyword latent semantic ranked search over encrypted cloud data and retrieve the most relevant files. We define a new scheme named Latent Semantic Analysis (LSA)-based multi-keyword ranked search which supports multi-keyword latent semantic ranked search. By using LSA, the proposed scheme could return not only the exact matching files, but also the files including the terms latent semantically associated to the query keyword. For example, when the user inputs the keyword "automobile" to search files, the proposed method returns not only the files containing "automobile", but also the files 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. To meet the challenge of supporting such multi-keyword semantic without privacy breaches, we propose the idea: the multi-keyword ranked search (MRSE) using "Latent Semantic Analysis".

PROBLEM FORMULATION:

A. SYSTEM MODEL:

The system model can be considered as three entities, as depicted in Figure 1: the data owner, the data user and the cloud server.

Data owner has a collection of data documents { , , ..., } d d d 1 2 m=D .A set of distinct keywords { , , ..., } w w w 1 2 n=W = 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 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. A corresponding trapdoor w T through search control mechanisms is generated. In this paper, we assume that the authorization between the data owner and the data user is approximately done.

Cloud server received w T from the authorized user. Then, the cloud server calculates and returns to the corresponding set

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

of encrypted documents. Moreover, to reduce the communication cost, the data user may send an optional number l along with the trapdoor T so that the cloud server only sends back top-l files that are most relevant to the search query.

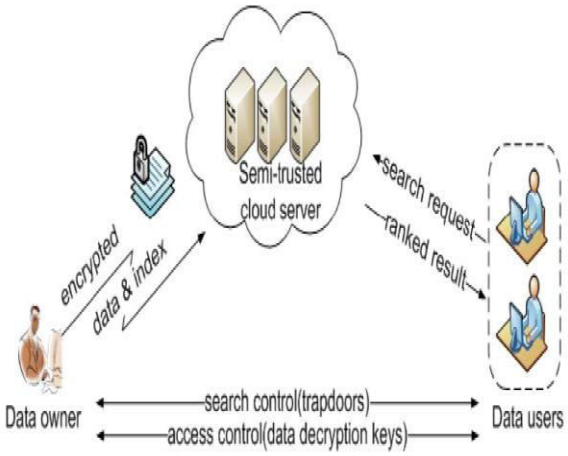

Fig: 1. Architecture of the search over encrypted cloud data.

B. 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 analyzes data in its storage and message flows received during the protocol so as to learn additional information. In this paper, we purpose to achieve security and ranked search under the above model. The designed goals of our system are following:

Latent Semantic Search: We aim to discover the latent semantic relationship between terms and documents. We use statistical techniques to estimate the latent semantic structure, and get rid of the obscuring "noise". The proposed scheme tries to put similar items near each other in some space in order that it could return the data user the files contain the terms latent semantically associated with the query keyword.

Multi-keyword Ranked Search: It supports both multi-keyword query and support result ranking.

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 could do some statistical analysis over the search result. Meanwhile, the same query should generate different trapdoors when searched twice. 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.

C.NOTATIONSANDPRELIMINARIES:

- D --the plaintext document collection, denoted as a set of n data documents { , , ..., } d_1 d_2 D_m
- C --the encrypted document collection stored in the cloud server, denoted as { , , ..., } c_1 c_2C_m
- W= --the dictionary, the keyword set composing of m keyword, denoted as { , , ..., } w_1 w_2w_m
- -W --the subset of W= ,

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

representing the keywords in a search request.

- I --the searchable index associated C , denoted as (, , ...,) $I_1 I_2 I_m$.
- Q --the query vector indicating the keywords of interest where each bit ∈Q [ ] {0,1} j represents the existence of the corresponding keyword in the query W.

PROPOSED SCHEME:

In this section, we give a detailed description of our scheme. We firstly propose to employ "Latent Semantic Analysis" to implement the latent semantic multi-keyword ranked search. A. Our Scheme Data owner wants to outsource m data files { , , ..., } d d d 1 2 m=D that he prepares to outsource to the cloud server in encrypted form while still keeping the capability to search through them. To do so, data owner firstly builds a secure searchable index I from a set of n distinct keywords W= extracted from the file collection D . According to the above definition about LSA, the data owner builds a term-document matrix ′A . Matrix ′A can be decomposed into the product of three other matrices. And then, 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 [16]. With t keywords of interest in W= as input, one binary vector Q is generated where each bit Q [ ]j indicates whether ∈W = W = j is true or false. The similarity score is expressed as the inner product of data vector A [ ]j and query vector Q .Specially, A [ ]j denotes the j -th column of the matrix A . λ S e tu p The data owner generates a 2 -bit vector as+n X and two + × +(2) (2) n n invertible matrices 1 2 { } M , M .The secret key SK is the form of a 3-tuple as 1 2 { } X , M , M . λ S K′B u ild In d ex(A ,) The data owner extracts a term-document matrix ′A from D and decomposes it into three other matrices U ×n t , S′ ×t t , V′ ×t m . According to our scheme, the data′ owner adopts statistical techniques to estimate the latent structure, and get rid of the obscuring "noise". To reduce dimensions, we choose previous k columns of ′S , and then deleting the corresponding columns of ′U and ′V respectively. Following, we multiply these three matrices ×U n k × , S k k ′ × , V k m′ ′ 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 -dimensions,) 2 + n (instead of n . Namely, the -th entry in) 1+ n (A [ ]j is set to a random number j ε , and the -th entry in) 2 + n (A [ ]j is set to 1 during the dimension extending. Finally, A [ ]j can be represented as (([ ]) , ,1) A T T j εj .The sub index 1 2 { [ ], [ ]} M A M A T T j · · =I j j ′′ ′ is built.

TRAPDOOR (W) With t keywords of interest in W as input, one binary vector Q is generated. The -th entry in) 1+ n (Q is set to a random number 1, and then scaled by a random number 0≠r , and the -th entry in) 2 + n (Q is set to another random number t during the dimension extending. Q can be represented as (Q ,) T T r r ,t . After applying the same splitting and encryption as above, the trapdoor w T is generated as $\{M_1^{-1} \cdot Q', M_2^{-1} \cdot Q''\}$.

QUERY (Tw ,l ,I) The inner product of j I and w T is calculated by the cloud server. After

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

sorting all scores, the cloud server returns the top-l ranked id list D w to the data user. The final similarity scores would be:

$$
\begin{aligned}
I_j \cdot T_w &= \{M_1^T \cdot A'[j], M_2^T \cdot A''[j]\} \cdot \{M_1^{-1} \cdot Q', M_2^{-1} \cdot Q''\} \\
&= (A'[j])^T \cdot Q' + (A''[j])^T \cdot Q'' \\
&= (A[j])^T \cdot Q \\
&= (A[j], \varepsilon_j, 1) \cdot (rQ^T, r, t)^T \\
&= r(A[j] \cdot Q^T + \varepsilon_j) + t
\end{aligned}
$$

Note that in our scheme, we add some random numbers to the final score, which clearly displays security strength.

PERFORMANCE AND SECURITY ANALYSIS: In this section, we show a thorough experimental evaluation of the proposed technique on a real dataset: the MED dataset [17] .The whole experiment is implemented by C++ language on a computer with Core 2.83GHz Processor, on Windows 7 system. For the proposed scheme, we will reduce to separate dimensions. The performance of our method is compared with the original MRSE scheme.

C. 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.

D. 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 instances that are relevant, while recall is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance. F-measure that combines precision and recall is the harmonic mean of precision and recall. Here, we adopt F-measure to weigh the result of our experiments.

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. We also investigate some further enhancements of our ranked search mechanism, including supporting more search semantics, i.e., TF _ IDF, and dynamic data operations. Thorough analysis investigating privacy and

efficiency guarantees of proposed schemes is given, and experiments on the real-world data set show our proposed schemes introduce low overhead on both computation and communication. In our future work, we will explore checking the integrity of the rank order in the search result assuming the cloud server is untrusted

REFERENCES:

[1] M. Armbrust, "A view of cloud computing", Communications of the ACM, vol. 53, no. 4, (2010), pp. 50-58.

[2] D. Boneh, "Public key encryption with keyword search", Advances in Cryptology-Eurocrypt 2004, Springer, (2004).

[3] R. Curtmola, "Searchable symmetric encryption: improved definitions and efficient constructions", Proceedings of the 13th ACM conference on Computer and communications security, ACM, (2006).

[4] D. X. Song, D. Wagner and A. Perrig, "Practical techniques for searches on encrypted data. in Security and Privacy", 2000. S&P 2000, Proceedings 2000 IEEE Symposium, IEEE, (2000). [5] C. Wang, "Secure ranked keyword search over encrypted cloud data", Distributed Computing Systems (ICDCS), 2010 IEEE 30th International Conference, IEEE, (2010).

[6] N. Cao, "Privacy-preserving multi-keyword ranked search over encrypted cloud data", INFOCOM, 2011 Proceedings IEEE, IEEE, (2011)..