Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 6, June 2015, pg.494 – 500

RESEARCH ARTICLE

CONFIDENTIALITY RETENTION IN KEYWORD-LIST RANKED SEARCH FOR ENCRYPTED CLOUD DATA

Syed Fayeq Jeelani (Department of CSE, GITM Haryana), syed.faaiq@gmail.com

Narender Singh (Department of CSE, GITM Haryana), er.narender.singh@gmail.com

Abstract: The advancement in cloud computing has motivated the data owners to outsource their data management systems from local sites to commercial public cloud for great flexibility and economic savings .For real privacy, user identity should remain hidden from CSP (Cloud service provider) and to protect privacy of data, data which is sensitive is to be encrypted before outsourcing. By considering the large number of data users, documents in the cloud, it is important for the search service to allow multi keyword query and provide result similarity ranking to meet the effective need of data retrieval search and not often differentiate the search results. In this system, we define and solve the challenging problem of confidentiality retention in keyword list ranked search over encrypted cloud data, and establish a set of strict privacy requirements for such a secure cloud data utilization system to be implemented in real. We first propose a basic idea for the Keyword list Ranked Search over Encrypted cloud data based on secure inner product computation and efficient similarity measure of coordinate matching, i.e., as many matches as possible, in order to capture the relevance of data documents to the search query, then we give two significantly improved Ranked search schemes to achieve various stringent privacy requirements in two different threat models. Assignment of anonymous ID to the user to provide more security to the data on cloud server is done. To improve the search experience of the data search service, further extension of the two schemes to support more search semantics is done.

Keywords- Cloud computing, searchable encryption, privacy preserving, keyword search, ranked search Anonymization

1. INTRODUCTION

Cloud computing is the computing in which large groups of remote servers are networked to allow centralized data storage and online access to computer services or resources. Clouds can be classified as public, private or hybrid. Cloud computing is the long dreamed vision of computing as a utility, where cloud customers can remotely store their data into the cloud so as to enjoy the on-demand high-quality applications and services from a shared pool of configurable computing resources. Its great flexibility and economic savings are motivating both individuals and enterprises to outsource their local complex data management system into the cloud. To protect data privacy and combat unsolicited accesses in the cloud

and beyond, sensitive data, for example, e-mails, personal health records, photo albums, tax documents, financial transactions, and so on, may have to be encrypted by data owners before outsourcing to the commercial public cloud this, however, obsoletes the traditional data utilization service based on plaintext keyword search. The trivial solution of downloading all the data and decrypting locally is clearly impractical, due to the huge amount of bandwidth cost in cloud scale systems. Thus, exploring privacy preserving and effective search service over encrypted cloud data is of paramount importance. Considering the potentially large number of on-demand data users and huge amount of outsourced data documents in the cloud, this problem is particularly challenging as it is extremely difficult to meet also the requirements of performance, system usability, and scalability. Cloud computing, a relatively new technology, has been gaining immense popularity over the last few years where user can rent software, hardware, infrastructure and computational recourse as per user basis. The cloud has three service models. In the Software as a Service (SaaS) model, the software or the applications are hosted over the cloud. These services are available to the customers based on the pay-as-per-use model. Google Apps and SALESFORCE are examples of this model. The Platform as a Service (PaaS) model provides a hosting environment for the client's application. Examples for PaaS model are Google App Engine and Amazon Web Services. The Infrastructure as a Service (IaaS) model lets the client to dynamically scale up or scale down the resources like processing power, storage capacity, network bandwidth etc. Example: Amazon EC2, Eucalyptus, etc.

2. LITERATURE SURVEY

Ruturaj Desai, et al. [9], proposes a new scheme to solve the problem of multi keyword search over encrypted data using trusted third party in cloud computing. User will encrypt their data locally. Before encrypting data, the index will be created. Trusted third party will use all these indexes to search data similar to the search query of user. Using these search results, cloud server will send encrypted document to the user**.** Unfortunately, data encryption, which restricts user's ability to perform keyword search and father demands the protection of keyword privacy, makes the traditional plaintext search methods fail for encrypted cloud data. Ranked search greatly improves system usability by normal matching files in ranked order regarding to certain relevance criteria. Ranked search can also elegantly eliminate unnecessary network traffic by sending back only the most relevant data, which is highly desirable in the "pay-as-you-use" cloud paradigm. For privacy protection, such ranking operation, however, should not leak any keyword related information. Besides, to improve search result accuracy as well as to enhance the user searching experience, it is also necessary for such ranking system to support multiple keyword search, as single keyword search often yields far too coarse results.

Shiba Sampat Kale, et al. [10], defined and solved the challenging problem of privacy-preserving multi-keyword ranked search over encrypted cloud data (MRSE), and establish a set of strict privacy requirements for such a secure cloud data utilization system to be implemented in real. First a basic idea has been proposed for the Multi-keyword Ranked Search over Encrypted cloud data (MRSE) based on secure inner product computation and efficient similarity measure of coordinate matching, i.e., as many matches as possible, in order to capture the relevance of data documents to the search query, then two significantly improved MRSE schemes has been given to achieve various stringent privacy requirements in two different threat models. Assignment of anonymous ID to the user to provide more security to the data on cloud server is done. To improve the search experience of the data search service, further extension of the two schemes to support more search semantics is done.

3. PROBLEM DEFINITION

Considering a cloud data hosting service involving three entities; the data owner, the data user, and the cloud server. The data owner has a collection of data documents F to be outsourced to the cloud server in the encrypted form C. To enable the searching capability over C for effective data utilization, the data owner, before outsourcing, will first build an encrypted searchable index I from F, and then outsource both the index I and the encrypted document collection C to the cloud server. To search the document collection for t given keywords, an authorized user acquires a corresponding trapdoor T through search control mechanisms, for example, broadcast encryption. Upon receiving T from a data user, the cloud server is responsible to search the index I and return the corresponding set of encrypted documents. To improve the document retrieval accuracy, the search result should be ranked by the cloud server according to some ranking criteria (e.g., coordinate matching, as will be introduced shortly). Moreover, to reduce the communication cost, the data user may send an optional number k along with the trapdoor T so that the cloud server only sends back top-k documents that are most relevant to the search query. Finally, the access control mechanism is employed to manage decryption capabilities given to users and the data collection can be updated in terms of inserting new documents, updating existing documents, and deleting existing documents.

3.1 EXISTING SYSTEM

The Effective data retrieval need, the large amount of documents demand the cloud server to perform result relevance ranking, instead of returning undifferentiated results. Such ranked search system enables data users to find the most relevant information quickly, rather than burdensomely sorting through every match in the content collection. Ranked Search also elegantly eliminate unnecessary network traffic by sending back only the most relevant data, which is highly desirable in the "pay-as-you-use" cloud paradigm. Thus, exploring privacy preserving and effective search service over encrypted cloud data is of paramount importance.

Considering the potentially large number of on-demand data users and huge amount of outsourced data documents in the cloud, this problem is particularly challenging as it is extremely difficult to meet also the requirements of performance, system usability, and scalability. The encrypted cloud data search system remains a very challenging task because of inherent security and privacy obstacles, including various strict requirements. They are still not adequate to provide users with acceptable result ranking functionality.

3.2 PROPOSED SYSTEM

For the first time, we define and solve the problem of keyword list ranked search over encrypted cloud data while preserving strict system wise privacy in the cloud computing paradigm. Among various multi-keyword semantics, we choose the efficient similarity measure of co-ordinate matching, i.e. as many matches as possible, to capture the relevance of data documents to the search query. Specifically we use inner product similarity i.e. the number of query keywords appearing in a document, to quantitatively evaluate such similarity measure of that document to the search query. Search results should be ranked by the cloud server according to some ranking criteria to reduce the communication cost.

4. IMPLEMENTATION

4.1 MRSE FRAMEWORK

With focus on the index and query, the MRSE system consists of four algorithms as follows:

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

Build Index (F, SK). Based on the data set 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 (W). With t keywords of interest in W as input, this algorithm generates a corresponding trapdoorT_W.

Query (T_W, k, I). When the cloud server receives a query request as (T_W, k),it performs the ranked search on the index I with the help of trapdoor T_W, and finally returns F_W, the ranked id list of top-k documents sorted by their similarity with W.

4.2 CONFIDENTIALITY PRESERVING AND EFFICIENT MRSE

To efficiently achieve multi-keyword ranked search, we propose to employ "inner product similarity" to quantitatively evaluate the efficient similarity measure "coordinate matching." Specifically, $D_i$is a binary data vector for document F_i where each bit D_i[j] € {0,1}represents the existence of the corresponding keyword W_j in that document, and Q is a binary query vector indicating the keywords of interest where each bit Q[j] € {0,1} represents the existence of the corresponding keyword W_j in the query W. The similarity score of document F_i to query W is therefore expressed as the inner product of their binary column vectors, i.e., D_i.Q. For the purpose of ranking, the cloud server must be given the capability to compare the similarity of different documents to the query. But, to preserve strict system wise privacy, data vector D_i, query vector Q and their inner product D_i.Q should not be exposed to the cloud server. We first propose a basic idea for the MRSE using secure inner product computation, which is adapted from a secure kNN technique, and then show how to significantly improve it to be privacy-preserving against different threat models in the MRSE framework.

4.3 MRSE_I_TF

In the ranking principle "coordinate matching," the presence of keyword in the document or the query is shown in the data vector or the query vector. Actually, there are more factors which could make impact on the search usability. For example, when one keyword appears in most documents in the data set, the importance of this keyword in the query is less than other keywords which appears in fewer documents. Similarly, if one document contains a query keyword in multiple locations, the user may prefer this to the other document which contains the query keyword in only one location. To capture these information in the search process, we use the TF × IDF weighting rule within the vector space model to calculate the similarity, where TF (or term frequency) is the number of times a given term or keyword appears within a file (to measure the importance of the term within the particular file), and IDF (or inverse document frequency) is obtained by dividing the number of files in the whole collection by the number of files containing the term (to measure the overall importance of the term within the whole collection). Among several hundred variations of the TF × IDF weighting scheme, no single combination of them outperforms any of the others universally.

Therefore, the similarity of the document and the query in terms of the cosine of the angle between the document vector and the query vector could be evaluated by computing the inner product of sub-index $I_i$and trapdoor T_w. Although this similarity measurement introduces more computation cost during the index construction and trapdoor generation, it captures more related information on the content of documents and query which returns better results of users' interest.

4.4 SUPPORTING DATA DYNAMICS

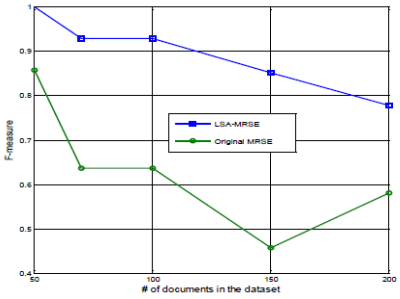
After the data set is outsourced to the cloud server, it may be updated in addition to being retrieved. Along with the updating operation on data documents, supporting the score dynamics in the searchable index is thus of

practical importance. While we consider three dynamic data operations as inserting new documents, modifying existing documents, and deleting existing documents, corresponding operations on the searchable index includes generating new index, updating existing index, and deleting existing index. Since dynamic data operations also affect the document frequency of corresponding keywords, we also need to update the dictionary W

5. RESULTS AND ANALYSIS

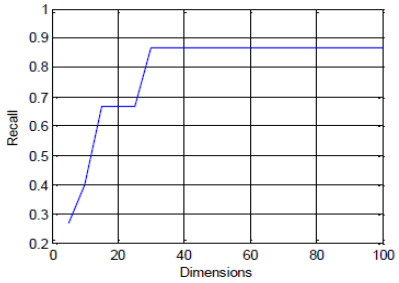
5.1 PERFORMANCE ANALYSIS

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.



Comparison of two Schemes

However, our scheme can make up for this disadvantage, and retrieve the most relevant files. Graph 7.1 shows that our method achieves remarkable result.



Recall of separate dimensions

Compared to the traditional vector space, smaller the latent semantic spaces, clearer the semantic relationships. Yet, the fact is that the lower dimension will not bring the better result. For example, we will use the 100 documents of MED to do the test and reduce separate dimensions respectively. Figure 7.2 shows a recall-dimension curve. From the Figure 7.2, the dimension reduces from 100 to 30, the recall has no change. It means that the relevant documents can be retrieved. Obviously, after the dimensions descended to 30, the values of the recall go down .It means that some relevant documents cannot be searched. Thus, when we conduct the experiments, we need to choose the appropriate dimension to achieve the best effect of experiment.

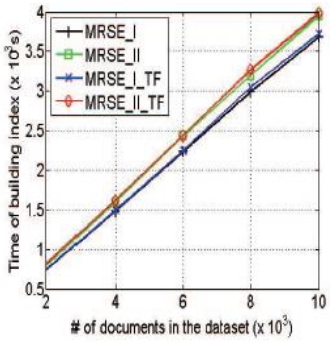
In our proposed scheme Index Confidentiality, Trapdoor Un-link ability and Keyword Privacy are all protected with remarkable improvements.

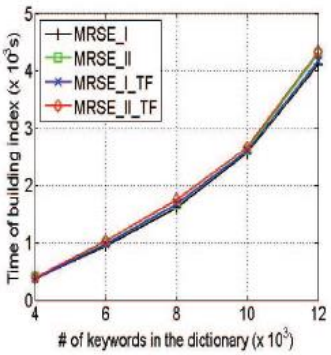
6. EFFICIENCY

Index Construction

To build a searchable sub-index $I_i$for each document F_i inthe data set F, the first step is to map the keyword set extracted from the document F_i to a data-vectorD_i, followed by encrypting every data vector. The time cost of mapping or encrypting depends directly on the dimensionality of data vector which is determined by the size of the dictionary, i.e., the number of indexed keywords.

Variable dataset, same dictionary, n = 4000. Same dataset, variable dictionary, m = 1000





And the time cost of building the whole index is also related to the number of sub-index which is equal to the number of documents in the data set. Figure 7.3 shows that, given the same dictionary where | W | = 4000, the time cost of building the whole index is nearly linear with the size of data set since the time cost of building each sub-index is fixed.

As shown in both figures additional computation in the TF × IDF weighting rule is insignificant considering much more computation are caused by the splitting process and matrix multiplication.. The size of sub-index is absolutely linear with the dimensionality of data vector which is determined by the number of keywords in the dictionary. The sizes of sub-index are very close in the two MRSE schemes because of trivial differences in the dimensionality of data vector.

Trapdoor Generation

Like index construction, every trapdoor generation incurs two multiplications of a matrix and a split query vector, where the dimensionality of matrix or query vector is different in two proposed schemes and becomes larger with the increasing size of dictionary. More importantly, it shows that the number of query keywords has little influence on the overhead of trapdoor generation which is a significant advantage over related works on multi-keyword searchable encryption.

Query

Query execution in the cloud server consists of computing and ranking similarity scores for all documents in the data set. The computation of similarity scores for the whole data collection is O(mn) in MRSE_I and MRSE_I_TF, and the computation increases to O(m(n+U)) in MRSE_II and MRSE_II_TF. The two schemes in the known cipher text model as MRSE_I and MRSE_I_TF have very similar query speed since they have the same dimensionality which is the major factor deciding the computation cost in the query. The query speed difference between MRSE_I and MRSE_I_TF or between MRSE_II and MRSE_II_TF is also caused by the dimensionality of data vector and query vector. With respect to the

communication cost in Query, the size of the trapdoor is the same as that of the sub-index, which keeps constant given the same dictionary, no matter how many keywords are contained in a query.

CONCLUSION AND FUTURE WORK

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 entrusted. As our future work, we will concentration the encrypted data of semantic keyword search in order that we can confront with more sophisticated search.

REFERENCES

[1] Ankatha Samuyelu Raja Vasanthi ," Secured Multi keyword Ranked Search over Encrypted Cloud Data", 2012
[2] 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.
[3] S. Kamara and K. Lauter, "Cryptographic Cloud Storage," *Proc. 14th Int'l Conf. Financial Cryptograpy and Data Security*, Jan. 2010.
[4] Y. Prasanna, Ramesh . "Efficient and Secure Multi-Keyword Search on Encrypted Cloud Data", 2012.
[5] Jain Wang, Yan Zhao , Shuo Jaing, and Jaijin Le, "Providing Privacy Preserving in Cloud Computing",2010.
[6] Larry A. Dunning, Ray Kresman ," Privacy Preserving Data Sharing With Anonymous ID Assignment",2013.
[7] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, "Fuzzy Keyword Search Over Encrypted Data in Cloud Computing," *Proc. IEEE INFOCOM,* Mar. 2010.
[8] 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. *IEEE INFOCOM,* 2010.
[10] C. Wang, Q. Wang, K. Ren, and W. Lou, "Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing," *Proc. IEEE INFOCOM,* 2010.
[11] N. Cao, Z. Yang, C. Wang, K. Ren, and W. Lou, "Privacy preserving Query over Encrypted Graph-Structured Data in Cloud Computing," *Proc. Distributed Computing Systems (ICDCS*), pp. 393-402, June, 2011.
[12] 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.
[13] S. Kamara and K. Lauter, "Cryptographic cloud storage," in *RLCPS, January 2010, LNCS. Springer, Heidelberg*.
[14] A. Singhal, "Modern information retrieval: A brief overview," *IEEE Data Engineering Bulletin*, vol. 24, no. 4, pp. 35–43, 2001.
[15] I. H. Witten, A. Moffat, and T. C. Bell, "Managing gigabytes: Compressing and indexing documents and images," Morgan Kaufmann Publishing, San Francisco, May 1999.

