IJDCST @Oct-Nov-2015, Issue- V-3, I-7, SW-18
ISSN-2320-7884 (Online)
ISSN-2321-0257 (Print)

Performing Preserving Privacy Multi Keyword Ranked Search on Encrypted Cloud Data

Dileep Nidamanuri[1], MD Amanatulla[2]
[1]M.Tech (CSE), Nimra Institute of Science and Technology, A.P., India.
[2]Asst Professor, Dept. of Computer Science & Engineering, Nimra Institute of Science and Technology, A.P., India.

Abstract — Inspired and motivated by the widespread development and deployment of public cloud computing infrastructures, using clouds to host data query services has become an fascinating solution for the advantages on scalability and cost-saving. But for protecting data privacy, sensitive data have to be encrypted before outsourcing, which obsoletes traditional data utilization based on plaintext keyword search. Thus, enabling an encrypted cloud data search service is of paramount importance. Considering the large number of data users and documents in the cloud, it is necessary to allow multiple keywords in the search request and return documents in the order of their relevance to these keywords. Related works on searchable encryption focus on single keyword search or Boolean keyword search, and rarely sort the search results. With the advent of cloud computing, data owners are motivated to outsource their complex data management systems from local sites to the commercial public cloud for great flexibility and economic savings. But for protecting data privacy, sensitive data have to be encrypted before outsourcing, which obsoletes traditional data utilization based on plaintext keyword search. 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. In this paper, for the first time, we define and solve the challenging problem of privacy-preserving multi-keyword ranked search over encrypted data in cloud computing (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. We first propose a basic idea for the MRSE based on secure inner product computation, and then give two significantly improved MRSE schemes to achieve various stringent privacy requirements in two different threat models. To improve search experience of the data search service, we further extend these two schemes to support more search semantics. Thorough analysis investigating privacy and efficiency guarantees of proposed schemes is given. Experiments on the real-world data set further show proposed schemes indeed introduce low overhead on computation and communication.

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

I. INTRODUCTION

Serving utilities on a service basis is a long dreamed vision of computing, 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 [2], [3]. 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 [4]; 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. Moreover, aside from eliminating the local storage management, storing data into the cloud serves no purpose unless they can be easily searched and utilized. 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.

On the one hand, to meet 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 [5]. 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. On the other hand, to improve the search result accuracy as well as to enhance the user searching experience, it is also necessary for such ranking system to support multiple keywords search, as single keyword search often yields far too coarse results. As a common practice indicated by today's web search engines (e.g., Google search), data users may tend to provide a set of keywords instead of only one as the indicator of their search interest to retrieve the most relevant data. And each keyword in the search request is able to help narrow down the search result further. "Coordinate matching" [6], i.e., as many matches as possible, is an efficient similarity measure among such multi-keyword semantics to refine the result relevance, and has been widely used in the plaintext information retrieval (IR) community. However, how to apply it in the encrypted cloud data search system remains a very challenging task because of inherent security and privacy obstacles, including various strict requirements like the data privacy, the index privacy, the keyword privacy, and many others.

In the literature, searchable encryption [7] is a helpful technique that treats encrypted data as documents and allows a user to securely search through a single keyword and retrieve documents of interest. However, direct application of these approaches to the secure large scale cloud data utilization system would not be necessarily suitable, as they are developed as crypto primitives and cannot accommodate such high service-level requirements like system usability, user searching experience, and easy information discovery. Although some recent designs have been proposed to support Boolean keyword search [8] as an attempt to enrich the search flexibility, they are still not adequate to provide users with acceptable result ranking functionality (see Section 7). Our early works have been aware of this problem, and provide solutions to the secure ranked search over encrypted data problem but only for queries consisting of a single keyword. How to design an efficient encrypted data search mechanism that supports multi-keyword semantics without privacy breaches still remains a challenging open problem.

In this paper, for the first time, we define and solve the problem of multi-keyword ranked search over encrypted cloud data (MRSE) while preserving strict system wise privacy in the cloud computing paradigm. 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. Specifically, we use "inner product similarity" [6], i.e., the number of query keywords appearing in a document, to quantitatively evaluate such similarity measure of that document to the search query. During the index construction, each document is associated with a binary vector as a sub-index where each bit represents whether corresponding keyword is contained in the document. The search query is also described as a binary vector where each bit means whether corresponding keyword appears in this search request, so the similarity could be exactly measured by the inner product of the query vector with the data vector. However, directly outsourcing the data vector or the query vector will violate the index privacy or the search privacy. To meet the challenge of supporting such multi-keyword semantic without privacy breaches, we propose a basic idea for the MRSE using secure inner product computation, which is adapted from a secure k-nearest neighbor (kNN) technique, and then give two significantly improved MRSE schemes in a step-by-step manner to achieve various stringent privacy requirements in two threat models with increased attack capabilities. Our contributions are summarized as follows:

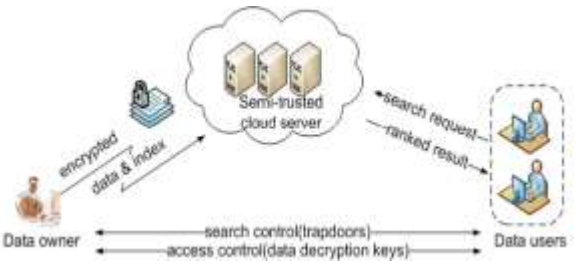

Figure 1 Architecture of the Search over encrypted cloud data

1. For the first time, we explore the problem of multikeyword ranked search over encrypted cloud data, and establish a set of strict privacy requirements for such a secure cloud data utilization system.

2. We propose two MRSE schemes based on the similarity measure of "coordinate matching" while meeting different privacy requirements in two different threat models.

3. We investigate some further enhancements of our ranked search mechanism to support more search semantics and dynamic data operations.

4. Thorough analysis investigating privacy and efficiency guarantees of the proposed schemes is given, and experiments on the real-world data set further show the proposed schemes indeed introduce low overhead on computation and communication.

Compared with the preliminary version [1] of this paper, this journal version proposes two new

mechanisms to support more search semantics. This version also studies the support of data/index dynamics in the mechanism design. Moreover, we improve the experimental works by adding the analysis and evaluation of two new schemes. In addition to these improvements, we add more analysis on secure inner product and the privacy part.

II. PROBLEM FORMULATION

System Model

Considering a cloud data hosting service involving three different entities, as illustrated in Fig. 1: 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 [10]. 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.

Threat Model

The cloud server is considered as “honest-but-curious” in our model, which is consistent with related works on cloud security. Specifically, the 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 the cloud server knows, we consider two threat models with different attack capabilities as follows.

Known ciphertext model. In this model, the cloud server is supposed to only know encrypted data set C and searchable index I, both of which are outsourced from the data owner. Known background model. In this stronger model, the cloud server is supposed to possess more knowledge than what can be accessed in the known ciphertext model. Such information may include the correlation relationship of given search requests (trapdoors), as well as the data set related statistical information. As an instance of possible attacks in this case, the cloud server could use the known trapdoor information combined with document/keyword frequency to deduce/identify certain keywords in the query.

Design Goals

To enable ranked search for effective utilization of outsourced cloud data under the aforementioned model, our system design should simultaneously achieve security and performance guarantees as follows.

- Multi-keyword ranked search. To design search schemes which allow multi-keyword query and provide result similarity ranking for effective data retrieval, instead of returning undifferentiated results.
- Privacy-preserving. To prevent the cloud server from learning additional information from the data set and the index, and to meet privacy requirements specified.
- Efficiency. Above goals on functionality and privacy should be achieved with low communication and computation overhead.

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. Fig. 6 shows the query time is dominated by the number of documents in the data set while the number of keywords in the query has very slight impact on it like the cost of trapdoor generation above. The two schemes in the known ciphertext 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 listed in the Table 3, which keeps constant given the same dictionary, no matter how many keywords are contained in a query.

While the computation and communication cost in the query procedure is linear with the number of query keywords in other multiple-keyword search schemes, our proposed schemes introduce nearly constant overhead while increasing the number of query keywords. Therefore, our schemes cannot be compromised by timing-based side channel attacks that try to differentiate certain queries based on their query time.

III. RELATED WORK

Single Keyword Searchable Encryption

Traditional single keyword searchable encryption schemes [9], [10] usually build an encrypted searchable index such that its content is hidden to the server unless it is given appropriate trapdoors generated via secret key(s) [4]. It is first studied by Song et al. [7] in the symmetric key setting, and improvements and advanced security definitions are given in Goh [8], Chang et al. [9], and Curtmola et al. [10]. Our early works[12] solve secure ranked keyword search which utilizes keyword frequency to rank results instead of returning undifferentiated results. However, they only supports single keyword search. In the public key setting, Boneh et al. [11] present the first searchable encryption construction, where anyone with public key can write to the data stored on server but only authorized users with private key can search. Public key solutions are usually very computationally expensive however. Furthermore, the keyword privacy could not be protected in the public key setting since server could encrypt any keyword with public key and then use the received trapdoor to evaluate this ciphertext.

Boolean Keyword Searchable Encryption

To enrich search functionalities, conjunctive keyword search over encrypted data have been proposed. These schemes incur large overhead caused by their fundamental primitives, such as computation cost by bilinear map, for example, or communication cost by secret sharing, for example, [7]. As a more general search approach, predicate encryption schemes are recently proposed to support both conjunctive and disjunctive search. Conjunctive keyword search returns "all-or-nothing," which means it only returns those documents in which all the keywords specified by the search query appear; disjunctive keyword search returns undifferentiated results, which means it returns every document that contains a subset of the specific keywords, even only one keyword of interest. In short, none of existing Boolean keyword searchable encryption schemes support multiple keywords ranked search over encrypted cloud data while preserving privacy as we propose to explore in this paper. Note that, inner product queries in predicate encryption only predicates whether two vectors are orthogonal or not, i.e., the inner product value is concealed except when it equals zero. Without providing the capability to compare concealed inner products, predicate encryption is not qualified for performing ranked search. Furthermore, most of these schemes are built upon the expensive evaluation of pairing operations on elliptic curves. Such inefficiency disadvantage also limits their practical performance when deployed in the cloud. Our early work [1] has been aware of this problem, and provides solutions to the multi-keyword ranked search over encrypted data problem. In this paper, we extend and improve more technical details as compared to [1]. We propose two new schemes to support more search semantics which improve the search experience of the MRSE scheme, and also study the dynamic operation on the data set and index which addresses some important yet practical considerations for the MRSE design. On a different front, the research on top-k retrieval in database community is also loosely connected to our problem. Besides, Cao et. al. proposed a privacy-preserving graph containment query scheme which solves the search problem with graph semantics.

IV. 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] 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, Apr, 2011.
[2] 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.

[3] 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.
[4] S. Kamara and K. Lauter, "Cryptographic Cloud Storage," Proc. 14th Int'l Conf. Financial Cryptograpy and Data Security, Jan. 2010.
[5] A. Singhal, "Modern Information Retrieval: A Brief Overview," IEEE Data Eng. Bull., vol. 24, no. 4, pp. 35-43, Mar. 2001.
[6] I.H. Witten, A. Moffat, and T.C. Bell, Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann Publishing, May 1999.
[7] D. Song, D. Wagner, and A. Perrig, "Practical Techniques for Searches on Encrypted Data," Proc. IEEE Symp. Security and Privacy, 2000.
[8] E.-J. Goh, "Secure Indexes," Cryptology ePrint Archive, http:// eprint.iacr.org/2003/216. 2003.
[9] 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.
[10] R. Curtmola, J.A. Garay, S. Kamara, and R. Ostrovsky, "Searchable Symmetric Encryption: Improved Definitions and Efficient Constructions," Proc. 13th ACM Conf. Computer and Comm. Security (CCS '06), 2006.
[11] D. Boneh, G.D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public Key Encryption with Keyword Search," Proc. Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT), 2004.
[12] M. Bellare, A. Boldyreva, and A. ONeill, "Deterministic and Efficiently Searchable Encryption," Proc. 27th Ann. Int'l Cryptology Conf. Advances in Cryptology (CRYPTO '07), 2007.