ISSN: 2348-2079

MAY-2015

International Journal of Intellectual Advancements and Research in Engineering Computations

SEARCHING ON ENCRYPTED DOCUMENTS UNDER CLOUDS

[1]S. Keerthiga, [2]S. Savitha Karpagam.

ABSTRACT

Sensitive cloud data have to be encrypted to protect data privacy, before outsourced to the commercial public cloud. The encryption process makes effective data utilization service a very challenging task. Traditional searchable encryption techniques allow users to securely search over encrypted data through keywords. The privacy enabled data searching scheme provides solution for secure ranked keyword search over encrypted cloud data. Ranked search greatly enhances system usability by enabling search result relevance ranking instead of sending undifferentiated results and further ensures the file retrieval accuracy. The statistical measure approach, i.e., relevance score, from information retrieval is explored to build a secure searchable index. One-to-many order-preserving mapping technique is developed to properly protect those sensitive score information. The system facilitates server-side ranking without losing keyword privacy. The system is improved to support relevance score dynamics process. Search result authentication is also provided in the system. One-to-many order-preserving mapping technique is also enhanced in reversible manner. The similarity analysis scheme is used to identify the query results under the cloud data storage.

INTRODUCTION

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 [5]. The benefits brought by this new computing model include but are not limited to: relief of the burden for storage management, universal data access with independent geographical locations and avoidance of capital expenditure on hardware, software and personnel maintenances, etc., [3].

As Cloud Computing becomes prevalent, more and more sensitive information are being centralized into the cloud, such as e-mails, personal health records, company finance data and government documents, etc. The fact that data owners and cloud server are no longer in the same trusted domain may put the outsourced unencrypted data at risk [4] the cloud server may leak data information to unauthorized entities [10] or even be hacked [6]. It follows that sensitive data have to be encrypted prior to outsourcing for data privacy and combating unsolicited accesses. Data encryption makes effective data utilization a very challenging task given that there could be a large amount of outsourced data files. Besides, in Cloud Computing, data owners may share their outsourced data with a large number of users, who might want to only retrieve certain specific data files they are interested in during a given session. One of the most popular ways to do so is through keyword-based search. Such keyword search technique allows users to selectively retrieve files of interest and has been widely applied in plaintext search scenarios. Unfortunately, data encryption, which restricts user's ability to perform keyword search and further demands the protection of keyword privacy, makes the traditional plaintext search methods fail for encrypted cloud data.

Author for Correspondence:

[1]Final year ME CSE, Department of CSE, Velalar College of Engineering and Technology, Erode, Tamilnadu, India.
Mail Id: keerthiga099@gmail.com.

[2]Associate Professer, Department of CSE, Velalar College of Engineering and Technology, Erode, Tamilnadu, India.

Although traditional searchable encryption schemes allow a user to securely search over encrypted data through keywords without first decrypting it, these techniques support only conventional Boolean keyword search, without capturing any relevance of the files in the search result. When directly applied in large collaborative data outsourcing cloud environment, they may suffer from the following two main drawbacks. On the one hand, for each search request, users without pre-knowledge of the encrypted cloud data have to go through every retrieved file in order to find ones most matching their interest, which demands possibly large amount of post-processing overhead, On the other hand, invariably sending back all files solely based on presence/ absence of the keyword further incurs large unnecessary network traffic, which is absolutely undesirable in today's pay-as-you-use cloud paradigm. In short, lacking of effective mechanisms to ensure the file retrieval accuracy is a significant drawback of existing searchable encryption schemes in the context of Cloud Computing. Nonetheless, the state of the art in information retrieval (IR) community has already been utilizing various scoring mechanisms quantify and rank order the relevance of files in response to any given search query. Although the importance of ranked search has received attention for a long history in the context of plaintext searching by IR community, surprisingly, it is still being overlooked and remains to be addressed in the context of encrypted data search.

Therefore, how to enable a searchable encryption system with support of secure ranked search is the problem tackled in this paper. Our work is among the first few ones to explore ranked search over encrypted data in Cloud Computing. Ranked search greatly enhances system usability by returning the matching files in a ranked order regarding to certain relevance criteria, thus making one step closer toward practical deployment of privacy-preserving data hosting services in the context of Cloud Computing. To achieve our design goals on both system security and usability, we propose to bring together the advance of both crypto and IR community to design the ranked searchable symmetric encryption (RSSE) scheme, in the spirit of "as-strong-as-possible" security guarantee. Specifically, we explore the statistical measure approach from IR and text mining to embed weight information of each file during the establishment of searchable index before outsourcing the encrypted file collection. As directly outsourcing relevance scores will leak lots of sensitive frequency information against the keyword privacy, we then integrate a recent crypto primitive order-preserving symmetric encryption (OPSE) and properly modify it to develop a one-to-many order-preserving mapping technique for our purpose to protect those sensitive weight information, while providing efficient ranked search functionalities. Our contribution can be summarized as follows:

1. For the first time, we define the problem of secure ranked keyword search over encrypted cloud data and provide such an effective protocol, which fulfills the secure ranked search functionality with little relevance score information leakage against keyword privacy.
2. Thorough security analysis shows that our ranked searchable symmetric encryption scheme indeed enjoys "as-strong-as-possible" security guarantee compared to previous searchable symmetric encryption (SSE) schemes.
3. We investigate the practical considerations and enhancements of our ranked search mechanism, including the efficient support of relevance score dynamics, the authentication of ranked search results and the reversibility of our proposed one-to- many order-preserving mapping techniques.
4. Extensive experimental results demonstrate the effectiveness and efficiency of the proposed solution.

RELATED WORK

Traditional searchable encryption has been widely studied as a cryptographic primitive, with a focus on security definition formalizations and efficiency improvements. Song et al. first introduced the notion of searchable encryption. They proposed a scheme in the symmetric key setting, where each word in the file is encrypted independently under a special two-layered encryption construction. Thus, a searching overhead is linear to the whole file collection length. Goh developed a Bloom filter-based per-file index, reducing the workload for each search request proportional to the number of files in

the collection. Chang and Mitzenmacher also developed a similar per-file index scheme. To further enhance search efficiency, Curtmola et al. proposed a per-keyword-based approach, where a single encrypted hash table index is built for the entire file collection, with each entry consisting of the trapdoor of a keyword and an encrypted set of related file identifiers. Searchable encryption has also been considered in the public-key setting. Aiming at tolerance of both minor typos and format inconsistencies in the user search input, fuzzy keyword search over encrypted cloud data has been proposed by Li et al. in [9]. Very recently, a privacy-assured similarity search mechanism over outsourced cloud data has been explored by Wang et al. in [2]. Note that all these schemes support only Boolean keyword search and none of them support the ranked search problem which we are focusing on in this paper. Following our research on secure ranked search over encrypted data, very recently, Cao et al. [1] propose a privacy-preserving multi keyword ranked search scheme, which extends our previous work in [1] with support of multi keyword query. They choose the principle of "coordinate matching," i.e., as many matches as possible, to capture the similarity between a multi keyword search query and data documents and later quantitatively formalize the principle by a secure inner product computation mechanism. One disadvantage of the scheme is that cloud server has to linearly traverse the whole index of all the documents for each search request, while ours is as efficient as existing SSE schemes with only constant search cost on cloud server.

Secure top-k retrieval from Database Community from database community is the most related work to our proposed RSSE. The idea of uniformly distributing posting elements using an order-preserving cryptographic function. The order preserving mapping function proposed does not support score dynamics, i.e., any insertion and updates of the scores in the index will result in the posting list completely rebuilt. Zerr et al. use a different order-preserving mapping based on presampling and training of the relevance scores to be outsourced, which is not as efficient as our proposed schemes. Besides, when scores following different distributions need to be inserted, their score transformation function still needs to be rebuilt. On the contrary, in our scheme the score dynamics can be gracefully handled, which is an important benefit inherited from the original OPSE. This can be observed from the Binary Search (.).Where the same score will always be mapped to the same random-sized nonoverlapping bucket, given the same encryption key. In other words, the newly changed scores will not affect previous mapped values. We note that supporting score dynamics, which can save quite a lot of computation overhead when file collection changes, is a significant advantage in our scheme. Moreover, both works above do not exhibit thorough security analysis which we do in the paper.

Other related techniques. Allowing range queries over encrypted data in the public key settings, where advanced privacy-preserving schemes were proposed to allow more sophisticated multi attribute search over encrypted files while preserving the attributes' privacy. Though these two schemes provide provably strong security, they are generally not efficient in our settings, as for a single search request, a full scan and expensive computation over the whole encrypted scores corresponding to the keyword posting list are required. Moreover, the two schemes do not support the ordered result listing on the server side. Thus, they cannot be effectively utilized in our scheme since the user still does not know which retrieved files would be the most relevant.

PROBLEM STATEMENT

Sensitive cloud data have to be encrypted to protect data privacy, before outsourced to the commercial public cloud. The encryption process makes effective data utilization service a very challenging task. Traditional searchable encryption techniques allow users to securely search over encrypted data through keywords. Searchable encryption technique supports only Boolean search process. Large amount of users and data files are not efficiently handled by the searchable encryption model. The privacy enabled data searching scheme provides solution for secure ranked keyword search over encrypted cloud data. Ranked search enhances system usability by enabling search result relevance ranking. Relevance score is a statistical measure approach is used in information retrieval. Relevance score is used in secure searchable index preparation

process. One-to-many order-preserving mapping technique is used to properly protect those sensitive score information. The system facilitates server-side ranking without losing keyword privacy. Ranked Searchable Symmetric Encryption (RSSE) scheme is used to perform secured data retrieval process. The following drawbacks are identified in the existing system.

- Static relevance score model
- Complex reversible operation under order preserving scheme
- Result authentication is not provided
- Retrieval latency is high

SEARCHABLE ENCRYPTION SCHEME

In the introduction, we have motivated the ranked keyword search over encrypted data to achieve economies of scale for Cloud Computing. We start from the review of existing searchable symmetric encryption schemes and framework for our proposed ranked searchable symmetric encryption. Note that by following the same security guarantee of existing SSE, it would be very inefficient to support ranked search functionality over encrypted data, as demonstrated in our basic scheme. The discussion of its demerits will lead to our proposed scheme.

We follow the similar framework of previously proposed searchable symmetric encryption schemes and adapt the framework for our ranked searchable encryption system. A ranked searchable encryption scheme consists of four algorithms (KeyGen, BuildIndex, TrapdoorGen, Search Index). Our ranked searchable encryption system can be constructed from these four algorithms in two phases, Setup and Retrieval.

EFFICIENT RANKED SEARCHABLE SYMMETRIC ENCRYPTION SCHEME

The above straightforward approach demonstrates the core problem that causes the inefficiency of ranked searchable encryption. That is how to let server quickly perform the ranking without actually knowing the relevance scores. To effectively support ranked search over encrypted file collection, we now resort to the newly developed cryptographic primitive—order preserving symmetric encryption achieve more practical performance. Note that by resorting to OPSE, our security guarantee of RSSE is inherently weakened compared to SSE, as we now let server know the relevance order. This is the information we want to trade off for efficient RSSE. We will first briefly discuss the primitive of OPSE and its pros and cons [7]. Then, we show how we can adapt it to suit our purpose for ranked searchable encryption with an "as-strong-as-possible" security guarantee. Finally, we demonstrate how to choose different scheme parameters via concrete examples.

The OPSE is a deterministic encryption scheme where the numerical ordering of the plaintexts gets preserved by the encryption function. The first cryptographic study of OPSE primitive and provides a construction that is provably secure under the security framework of pseudorandom function or pseudorandom permutation. Namely, considering that any order-preserving function g (.) from domain D = {1, . . .,M} to range R = {1, . . .,N} can be uniquely defined by a combination of M out of N ordered items, an OPSE is then said to be secure if and only if an adversary has to perform a brute force search over all the possible combinations of M out of N to break the encryption scheme. If the security level is chosen to be 80 bits, then it is suggested to choose M = N/2 > 80 so that the total number of combinations will be greater than 2^{80}. Their construction is based on an uncovered relationship between a random order-preserving function and the hypergeometric probability distribution, which will later be denoted as HGD.

At the first glance, by changing the relevance score encryption from the standard indistinguishable symmetric encryption scheme to this OPSE, it seems to follow directly that efficient relevance score ranking can be achieved just like in the plaintext domain. As pointed out earlier, the OPSE is a deterministic encryption scheme. This inherent deterministic property, if not treated appropriately. One such information leakage is the plaintext distribution. For easy exposition, we encode the actual score into 128 levels in domain from 1 to 128.. In [8], the authors further point out that the TF distribution of the keyword in a given file collection

usually follows a power law distribution, regardless of the popularity of the keyword. Their results on a few test file collections show that not only different keywords can be differentiated by the slope and value range of their TF distribution. Thus, with certain background information on the file collection, such as knowing it contains only technical research papers, the adversary may be able to reverse engineer the keyword "network" directly from the encrypted score distribution without actually breaking the trapdoor construction, nor does the adversary need to break the OPSE.

SECURITY AND PRIVACY ENSURED DATA SEARCH MODEL FOR ENCRYPTED STORAGE

The cloud data center manages the transactional data values. The data values are maintained in encrypted format. The data values are queried using the encrypted query values. The system is designed to provide data security and privacy for the transactional data over the cloud environment. The order preserving mapping model is used for the encryption process. The score functions are used to fetch the data values in a ranked manner. The dynamic scoring mechanism is used in the system.

The system is divided into two applications. They are data source and client application. The data source manages the transactional data values. The client application issues the query value and collects the data from the data source. The data values are updated in the data source in an encrypted format. The data retrieval and ranking operations are carried out on the encrypted data format only. The system secures the data under the storage and query transmission process.

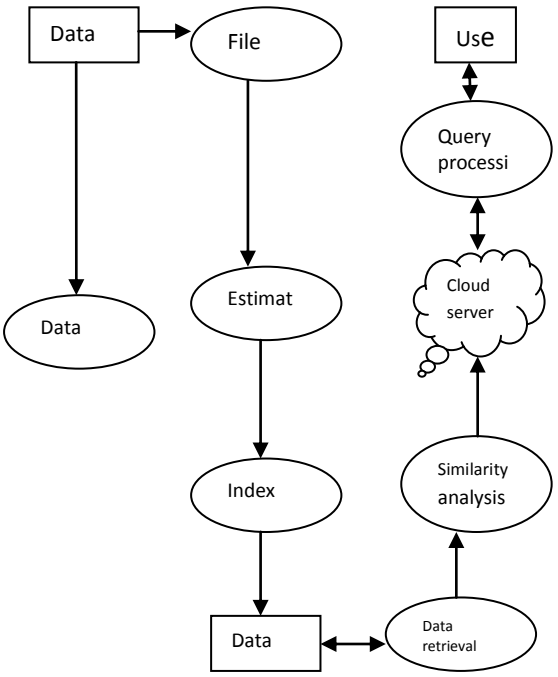

Fig.6.1: Encrypted Search Model

The system is divided into five major modules. They are data source, storage management, score assignment, client and query process. The data source module is designed to manage the data values. The storage management module is designed to perform the data encryption and update operations. The score

assignment module is used to assign the relevance score the for the transactional data values. The client application is used to fetch the data value from the data source. The query process module is designed to submit and collect the data values.

DATA SOURCE

The data source application is designed to manage the transactional and user information. The user information is updated with their access information. All the query history is maintained under the data source application. The transactional data values are maintained for different domains. The data retrieval is performed under the data source application.

STORAGE MANAGEMENT

The storage management is designed to handle data encryption and update operations. The order preserving mapping technique is used to encrypt the data values. The system includes the reversible order preserving map model for the encryption process. The data update operation can be dynamically performed on the system. The data values are updated and stored in the encrypted format. The transactional data and its encryption process are carried out under the data source environment.

SCORE ASSIGNMENT

The score assignment module is designed to assign the score values for the transactions. The similarity value is estimated to assign the score values. The relevance score is used to rank the transaction data values. The data retrieval is carried out with the score functions. The incremental data update initiates the dynamic score assignment process. The dynamic score assignment process updates the score values based on the new transaction data values.

CLIENT

The client application is designed to perform the data retrieval operations. The data values are collected from the server and updated into the client interface. Each client is authenticated with unique identification value. The client collects the data values with query keywords.

QUERY PROCESS

The query process module is designed to fetch the transactional data values. Query keyword is collected from the client. The query keyword is encrypted and transferred to the data source. The data source performs the searching process. The transactional data values are compared and similarity values are estimated. The results are prepared using the similarity value and threshold levels. The client application decrypts the transactional data values and produces the results in a ranked way.

CONCLUSION

Cloud customers can remotely store their data on a shared pool of configurable computing resources in cloud. Searchable Symmetric Encryption scheme is used to provide storage and retrieval security. Order Preserving Symmetric Encryption scheme is enhanced in reversible mechanism. The system is improved with result authentication and similarity based ranking model. The data storage and search process is carried out with encrypted query model. The system performs index operations on encrypted data values. The system also secures the search results. The system supports incremental data update scheme.

REFERENCES

[1]. N. Cao and W. Lou, "Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data," Proc. IEEE Infocom '11, 2011.

[2]. C. Wang, K. Ren, S. Yu, K. Mahendra and R. Urs, "Achieving Usable and Privacy-Assured Similarity Search over Outsourced Cloud Data," Proc. IEEE INFOCOM, 2012.

[3]. M. Armbrust, I. Stoica and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing," Technical Report UCB-EECS-2009-28, Univ. of California, Feb. 2009.

[4]. Cloud Security Alliance "Security Guidance for Critical Areas of Focus in Cloud Computing," http://www.cloudsecurityalliance.org, 2009.

[5]. C. Wang, N. Cao, J. Li and Lou, "Secure Ranked Keyword Search over Encrypted Cloud Data," Proc. IEEE 30th Int'l Conf. Distributed Computing Systems, 2010.

[6]. B. Krebs, "Payment Processor Breach May Be Largest Ever," http://voices.washingtonpost.com/securityfix/2009/01/payment_processor_breach_may_b.html, Jan. 2009.

[7]. Ning Cao, Ming Li, Kui Ren and Wenjing Lou, "Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data", IEEE Transactions On Parallel And Distributed Systems, Vol. 25, No. 1, January 2014

[8]. S. Zerr, D. Olmedilla, W. Nejdl and W. Siberski, "Zerber+r: Top-k Retrieval from a Confidential Index," Proc. Int'l Conf. Extending Database Technology: Advances in Database Technology (EDBT '09), 2009.

[9]. J. Li, Q. Wang, K. Ren and W. Lou, "Fuzzy Keyword Search over Encrypted Data in Cloud Computing," Proc. IEEE Infocom '10, 2010.

[10]. Z. Slocum, "Your Google Docs: Soon in Search Results?" http://news.cnet.com/8301-17939_109-10357137-2.html, 2009.