Efficient and Secure Multi-Keyword Search on Encrypted Cloud Data

Y. Prasanna, Ramesh
Dept. of CSE, Kakinada Institute of Engineering and Technology, Korangi, Kakinada, AP, India

Abstract

As Cloud Computing becomes prevalent, sensitive information are being increasingly centralized into the cloud. For the protection of data privacy, sensitive data has to be encrypted before outsourcing, which makes effective data utilization a very challenging task. Although traditional searchable encryption schemes allow users to securely search over encrypted data through keywords, these techniques support only boolean search, without capturing any relevance of data files. This approach suffers from two main drawbacks when directly applied in the context of Cloud Computing. On the one hand, users, who do not necessarily have pre-knowledge of the encrypted cloud data, have to post process every retrieved file in order to find ones most matching their interest; On the other hand, invariably retrieving all files containing the queried keyword further incurs unnecessary network traffic, which is absolutely undesirable in today’s pay-as-you-use cloud paradigm.

In this paper, for the first time we define and solve the problem of effective yet secure ranked keyword search over encrypted cloud data. Ranked search greatly enhances system usability by returning the matching files in a ranked order regarding to certain relevance criteria (e.g., keyword frequency), thus making one step closer towards practical deployment of privacy-preserving data hosting services in Cloud Computing. In this paper, for the first time, we define and solve 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 become a reality. Among various multi-keyword semantics, we choose 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. 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.

Keywords

I. Introduction

We are living in a highly networked environment, where huge amounts of data are stored in remote, but not necessarily trusted servers. There are several privacy issues regarding to accessing data on such servers; two of them can easily be identified: sensitivity of i) keywords sent in queries and ii) the data retrieved; both need to be hidden. A related protocol, Private Information Retrieval (PIR) [6] enables the user to access public or private databases without revealing which data he is extracting. Since privacy is of a great concern, PIR protocols have been extensively studied in the past [2, 6, 9, 10].

In today’s information technology landscape, customers that need high storage and computation power tend to outsource their data and services to clouds. Clouds enable customers to remotely store and access their data by lowering the cost of hardware ownership while providing robust and fast services [8]. The importance and necessity of privacy preserving search techniques are even more pronounced in the cloud applications. Due to the fact that large companies that operate the public clouds like Google or Amazon may access the sensitive data and search patterns, hiding the query and the retrieved data has great importance in ensuring the privacy and security of those using cloud services.

We aim to achieve an efficient system where any authorized user can perform a search on a remote database with multiple keywords, without revealing neither the keywords he searches for, nor the contents of the documents he retrieves. Our proposed system differs from the previous works which assume that only the data owner queries the database [2, 4]. In contrast to previous works, our proposal facilitate that a group of users can query the database provided that they possess trapdoors for search terms that authorize the users to include them in their queries. Moreover, our proposed system is able to perform multiple keyword search in a single query and ranks the results so the user can retrieve only the top matches.

The contributions of this paper can be summarized as follows. Firstly, we provide formal definitions for the security and privacy requirements of keyword search on encrypted cloud data. Secondly, we propose an efficient ranked multikey word search scheme and formally prove that it is secure in accordance with the defined requirements. Thirdly, we propose a ranking method that proves to be efficient to implement and effective in returning documents highly relevant to submitted search terms. Lastly, we implement the proposed scheme and demonstrate that it is much more efficient than existing methods in literature.

II. Related Work

The problem of Private Information Retrieval was first introduced by Chor et al. [6]. Recently Groth et al. [9] propose a multi-query PIR method with constant communication rate. However, any PIR-based technique requires highly costly cryptographic operations in order to hide the access pattern. This is inefficient in the large scale cloud system and as an alternative approach, privacy preserving search is employed which aims to hide the content of the retrieved data instead of which data is retrieved.

Ogata and Kurosawa [10] show privacy preserving keyword search protocol based on RSA blind signatures. The scheme requires a public-key operation per item in the database for every query and this operation must be performed on the user side. Freedman et al. [8], proposed an alternative implementation for private keyword search that uses homomorphic encryption and oblivious polynomial evaluation methods. The computation and communication costs of this method are quite large since every search term in a query requires several homomorphic encryption operations both on the server and the user side. A recent work proposed by Wang et al. [13] allows ranked search over an encrypted database by using inner product similarity. However, this work is only limited to
When a user wants to perform a keyword search, he first connects as its one of the steps. This process is referred as the index, it may learn the major subject of a document, even the means that the user must know a list of all valid keywords and their positions as a compulsory information to generate a query. This assumption may not be applicable in several cases. Moreover, it is not efficient due to matrix multiplication operations of square matrices where the number of rows is in the order of several thousands.

Wang et al. [14] propose a trapdoorless private keyword search scheme, where their model requires a trusted third party which they named as the Group Manager. We adapt indexing method to our scheme, but we use a totally different encryption methodology to increase the security and efficiency of the scheme.

![Architecture of Search Method](image)

**III. PROPOSED SYSTEM**

The problem that we consider is privacy-preserving keyword search on private database model, where the documents are simply encrypted with the secret keys unknown to the actual holder of the database (i.e. Cloud Server). We consider three roles coherent with previous works [3, 14]:

- Data owner, who is the actual owner of the database. The data owner collects and/or generates the information in the database and lacks the means (or is unwilling) to maintain/operate the database,
- Users are the members in a group who are entitled to access (part of) the information of the database,
- Server is a professional entity (e.g. cloud) to offer information services to authorized users. It is often required that the server is oblivious to content of the database it maintains, the search terms in queries and documents retrieved.

In fig. 1, steps and typical interactions between the participants of the system are illustrated. In an offline stage, the data owner creates a search index for each document. The search index file is created using a secret key based trapdoor generation function where the secret keys are only known by the data owner. Then, the data owner uploads these search index files to the server together with the encrypted documents. We use symmetric-key encryption as the encryption method since it can handle large document sizes efficiently. This process is referred as the index generation henceforth and the trapdoor generation is considered as its one of the steps.

When a user wants to perform a keyword search, he first connects to the data owner. He learns the trapdoors (cf. Step 1 in fig. 1) for the keywords he wants to search for, without revealing the keyword information to the data owner. Since the user can use the same trapdoor for many queries containing the corresponding search term, this operation does not need to be performed every time the user performs a query. After learning the trapdoor information, the user generates the query (referred as query generation henceforth) and submits it to the server (cf. step 2 in fig. 1). In return, he receives metadata2 for the matched documents in a rank ordered manner as will be explained in subsequent sections. Then the user retrieves the encrypted documents (cf. Step 3 in fig. 1) he chooses after analyzing the metadata that basically conveys a relevancy level of the each matched document, where the number of documents returned is specified by the user. Finally, the user interacts with the data owner in order to decrypt the documents and get the corresponding plaintext (cf. Step 4 in fig. 1); and in the process the data owner does not learn the documents.

**A. Privacy Requirements**

The privacy definition for search methods in the related literature is that the server should learn nothing but the search results [3]. We further tighten the privacy over this general privacy definition and establish a set of privacy requirements for privacy-preserving search protocols. A multi-keyword search method must provide the following user and data privacy properties (first intuitions and then formal definitions are given):

- (Data Privacy) No one but the user can learn the actual retrieved data.
- (Index Privacy) The search index or the query index do not leak any information about the corresponding keywords.
- (Trapdoor Privacy) Given one trapdoor for a set of keywords, the server cannot generate another valid trapdoor.
- (Non-Impersonation) No one can impersonate a legitimate user.

MRSE (multi-keyword ranked search over encrypted cloud data) consists of four algorithms as follows.

- Setup (1’’) Taking a security parameter ` as input, data owner outputs a symmetric key as SK.
- Build Index (F, SK) Based on the dataset F, data owner builds a searchable index I which is encrypted by the symmetric key SK and then outsourced to cloud server. After the index construction, the document collection can be independently encrypted and outsourced.
- Trapdoor(W) With keywords of interest in W as input, this algorithm generates a corresponding trapdoor T W
- Query(T W, k, l) When cloud server receives a query request as (T e W, k), it performs the ranked search on the index I with the help of trapdoor TW, and finally returns FW, the ranked id list of top-k documents sorted by their similarity with IW.

**B. Privacy Requirements for MRSE**

The representative privacy guarantee in the related literature, such as searchable encryption, is that the server should learn nothing but search results. With this general privacy description, we explore and establish a set of strict privacy requirements specifically for the MRSE framework.

As for the data privacy, data owner can resort to traditional symmetric key cryptography to encrypt the data before outsourcing, and successfully prevent cloud server from prying into outsourced data. With respect to the index privacy, if server deduces any association between keywords and encrypted documents from index, it may learn the major subject of a document, even the
content of a short document [23]. Therefore, searchable index should be constructed to prevent server from performing such kind of association attack. While data and index privacy guarantees are demanded by default in the related literature, various search privacy requirements involved in the query procedure are more complex and difficult to tackle as follows:

Keyword Privacy As users usually prefer to keep their search from being exposed to others like cloud server, the most important concern is to hide what they are searching, i.e., the keywords indicated by the corresponding trapdoor. Although the trapdoor can be generated in a cryptographic way to protect the query keywords, cloud server could do some statistical analysis over the search result to make an estimate. As a kind of statistical information, document frequency (i.e., the number of documents containing the keyword) is sufficient to identify the keyword with high probability [24]. When cloud server knows some background information of the dataset, this keyword specific information may be utilized to reverse engineer the keyword.

Trapdoor Privacy Since only authorized users are allowed to acquire trapdoors for their search query, the server is not expected to have the ability to generate valid trapdoors from previous received ones. Specifically, given one trapdoor for a set of multiple keywords, the server is not allowed to generate a valid trapdoor for its subset, including single keyword. For example, it is forbidden to generate or deduce a new trapdoor as TWi for keyword Wi from the received trapdoor as T(Wi;Wk) for two keywords (W, Wj). Moreover, the server is not allowed to generate a valid trapdoor, e.g., T(Wi;Wj ), from two or more trapdoors, like T(W,Wk) and T(W,Wk).

Search Pattern In accordance with the definition in related work on single keyword searchable encryption [8], search pattern of data user in MRSE means any information that can be derived by server if it acquires the knowledge that two arbitrary searches are performed for the same keywords or not. If the trapdoor is generated in a deterministic manner, server could easily know the search pattern of any data user by comparing trapdoors received from that user. So the fundamental protection for search pattern is to introduce non-determinacy into trapdoor generation procedure. Access Pattern Within the ranked search, access pattern is the sequence of search results where every search result is a set of documents with rank order. Specifically, the search result for FW is denoted as FW, consisting of the id list of all documents ranked by their similarity to FW. Then the access pattern is denoted as (FW1,FW2) which are the results of sequential searches. Although a few searchable encryption works, e.g., [17] has been proposed to utilize private information retrieval (PIR) technique [25] to hide access pattern, our proposed schemes are not designed to protect access pattern for the efficiency concerns. This is because any PIR based technique must “touch” the whole dataset outsourced on the server which is inefficient in the large scale cloud system.

IV. Conclusion

In this paper, we motivate and solve the problem of efficient and secure ranked multi-keyword search on remotely stored encrypted database model where the database users are protected against privacy violations. We first define the security requirements for the given problem. We then employ a secure usage of the method given in [14] for practical application scenarios where total number of keywords that can be searched is relatively limited and there are only few search terms in a query by using a trapdoor based system where the trapdoor can only be generated by the data owner. We appropriately increase the efficiency of the scheme by using symmetric-key encryption method rather than public-key encryption for document encryption. We also propose to use the blinded encryption technique in accessing the contents of the retrieved documents without revealing them to other parties. We prove that our proposed method satisfies the security requirements. The proposed ranking method proves to be efficient to return highly relevant documents corresponding to submitted search terms. We implement the entire scheme and extensive experimental results on the implementation demonstrate the effectiveness and efficiency of our solution.

References


Y. Prasanna is a student of Kakinada Institute of Engineering & Technology, Korangi, and Kakinada. Presently he is pursuing his M.TECH from this college and He received his graduation from JNTUK, in the year 2009.

Mr. Ramesh received the B. Tech degree in Computer Science and Engineering from JNTU, Hyderabad, Andhra Pradesh, India, in 2007, and also, received M.Tech, from GGET college, Rajamundry from JNTU, Hyderabad, Andhra Pradesh, India, in2008. He is working as Associate Professor in KIET, Kakinada. His research interests include Data Mining and Emotion Recognition and Speech Recognition in Image Processing. He is member of Computer Society of India and Indian Society for Technical Education.