International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

An Efficient Search in Cloud Computing Using Ranked Keyword Search Algorithm

R. Aravind

UG student, Department of Computer Science and Engineering, Saveetha School of Engineering, Thandalam, India

Abstract: ***Cloud computing is a subscription-based service where the networked storage space and computer resources can be obtained. Cloud computing economically enables the paradigm of data service outsourcing. However, to protect data privacy, sensitive cloud data have to be encrypted before outsourced to the commercial public cloud, which makes effective data utilization. In the proposed system, the problem of effective secure ranked keyword search over encrypted cloud data is done. Ranked keyword search greatly enhances the system usability by returning the matching files in a ranked order. The files are matched according to certain criteria and it makes one step closer towards the deployment of privacy-preserving data hosting services in Cloud Computing. The resulting design is able to facilitate efficient server-side ranking without losing keyword privacy. In the proposed system, the querying process over the cloud computing infrastructure (IAAS) using secured and encrypted data access and ranking is performed for the getting better results.***

Keywords: Ranked keyword search, confidential data, searchable encryption, cloud computing.

1. Introduction

Cloud Computing enables cloud customers to store their data into the cloud and provides an on-demand high quality applications and services from a shared pool of configurable computing resources. 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.

To protect data privacy, sensitive data has to be encrypted before outsourcing so as to provide end-to-end data confidentiality assurance in the cloud and beyond. Thus, exploring privacy-preserving and effective search service over encrypted cloud data is of paramount importance. Data owners may share their data with large number of on-demand data users and huge amount of outsourced data documents in 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, large amount of documents demand cloud server to perform result relevance ranking, instead of returning undifferentiated result. Such ranked search system enables data users to find the most relevant information quickly. Ranked search can also elegantly eliminate unnecessary network traffic by sending back only the most relevant data.

The other hand, to improve search result accuracy as well as enhance user searching experience, it is also crucial for such ranking system to support multiple keywords search, as single keyword search often yields far too coarse result. 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.

Some recent designs have been proposed to support Boolean keyword search as an attempt to enrich the search flexibility, they are still not adequate to provide users with acceptable result ranking functionality and solves the secure ranked search over encrypted data with support of only single keyword query.

The rest of the paper is organized as follows. In the next section, we present some of the related work in this direction.
Section III describes the system architecture for multi-keyword search. In section IV, the details about the system methodology is discussed. Section V outlines the future work and section VI concludes the paper.

2. Related Work

The evolution of cloud computing is one of the major advances in the technologies which represent cloud computing: Platform-as-a-service (PaaS), Software-as-a-Service (SaaS) and Infrastructure-as-a-Service (IaaS). Public key encryption [8] deals with the privacy of database data. There are two different scenarios: public databases and private databases. Private databases: A user wishes to upload its private data to a remote database and wishes to keep the data private from the remote database administrator. An additional privacy requirement is to hide any information from the database administrator regarding the access pattern, i.e. if some item was retrieved more than once, some item was not retrieved.

Public Databases: The database data is public (such as stock quotes) but the user is unaware of it and wishes to retrieve some data-item or search for some data-item, without revealing to the database administrator which item it is.

All keyword searches are based on this index; hence our scheme does not order full pattern-matching generality with the actual text. In practice, this should be sufficient for most users. It is worth noting that this framework can have complete control over what words are keywords that can be useful for many applications.

In Confidentiality-Preserving Rank-Ordered Search, when an authorized user remotely accesses the data to search and retrieve desired documents, the large size of the collections

often makes it infeasible to ship all encrypted data to the user's side, and then perform decryption and search on the user's trusted computers. Therefore, new techniques are needed to encrypt and organize the data collections in a way as to allow the data centre to perform efficient search in an encrypted domain [4].

Order-Preserving Symmetric Encryption (OPSE), is a deterministic encryption scheme whose encryption function preserves numerical ordering of the plaintexts. OPSE is the form of one-part codes, which are lists of plaintexts and the corresponding cipher texts, both of which are arranged in an alphabetical or numerical order so that a single copy is required for efficient encryption and decryption [5].

OPSE not only allows efficient range queries, but also allows indexing and query processing to be done exactly and is efficient for unencrypted data. Data owner has a collection of the data files and they outsource the data on to the cloud server in an encrypted form for the effective utilization of the data [1]. To do so, before outsourcing, data owner will first build a secure searchable index from a set of distinct keywords extracted from the file collection, and store both the index and the encrypted file collection on to the cloud server. To search the file collection for a given keyword, an authorized user generates and submits a search request in a secret form. An authorized user remotely accesses the data to search and retrieve the desired documents, which often makes it infeasible to ship all encrypted data to the user's side, and then perform decryption and search on the user's trusted computers. Therefore, new techniques are needed to encrypt and organize the data collections in such a way so as to allow the data centre to perform efficient search in encrypted domain [6]. The requirements of balancing privacy and confidentiality with efficiency and accuracy pose significant challenges to the design of search schemes for a number of search scenarios.

3. System Architecture

The overview of the proposed system is illustrated in Figure 1. We assume that the parties are semi-honest and do not collude with each other to bypass the security measures. The data owner uploads these search index files to the server together with the encrypted documents.

Figure 1: Multi-keyword Ranked Seacrh Architecture

The search index is created using a secret key based trapdoor generation function where the secret keys are only known by the data owner. The encryption method can handle large document size efficiently. Data owner can upload any text files on to the server, the main server will verify the index information present in it and diverts the query to the corresponding cloud servers. The main keywords are extracted and the unwanted words are filtered by stemming process. The file names are updated in the corresponding cloud servers. The query of the user is encrypted using RSA algorithm; this encryption process will prevent the data theft from the hackers. The files are retrieved to the user as the index data of all the files are maintained in the index of the main cloud server. Upon receiving data, cloud server is responsible to search the index and return the corresponding set of encrypted documents. To improve document retrieval accuracy, search result should be ranked by cloud server according to some ranking criteria (e.g., coordinate matching, as will be introduced shortly). Moreover, to reduce communication cost, data user may send an optional number so that cloud server only sends back top-*k* documents that are most relevant to the search query.

When a user wants to perform a keyword search, he first connects to the data owner and search for data without revealing the keyword information to the data owner. The user generates the query and submits it to the server. In return, he receives metadata for the matched documents in a rank ordered manner. Then the user retrieves the encrypted data he chooses after analysing 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. This is the process of multi-keyword search architecture.

4. System Methodology

Cloud owners publish certain files and they are encrypted so as to maintain privacy. The server can hold certain keywords and the process can be performed when any user accesses a particular data or file. This can be done based on ranking process. The scheme is shown as follows.

1. Setup: Data owner randomly generates a set of files on to the cloud.
2. Build Index: The cloud can be of index server which holds information about the servers.
3. Trapdoor: This can be preventing various attackers from accessing private files using encryption techniques.
4. Query: This can be given by the users to search for any particular file or information

4.1 Cloud Organization and Stemming

Cloud servers are constructed with the files and the index information are maintained in the main cloud server. Query is given to the main cloud server, so that the main cloud server will verify the index information present in it and divert the query to the corresponding cloud servers. The words in the files are extracted to filter the unwanted words using word stemmer algorithm.

The keywords are passed on to the cloud. The file names are updated in the corresponding cloud servers. The index server is the main server which holds all the information of the corresponding servers.

Cloud server intentionally wants to do so for saving cost when handling large number of search requests, or there may be software bugs, or internal/external attacks. Thus, enabling a search result authentication mechanism that can detect such unexpected behaviour of cloud server is also of practical interest and worth further investigation.

4.2 Encryption Technique

The query of the user is encrypted using RSA algorithm; this encryption process will prevent the data theft from the hackers. Data security is ensured using RSA encryption. RSA (which stands for Rivest, Shamir and Adleman who first publicly described it) is an algorithm for public-key cryptography.

The encrypted scores are the only additional information that the adversary can utilize against the security guarantee, i.e., keyword privacy and file confidentiality. 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.

Due to the security strength of the file encryption scheme, the file content is clearly well protected. Thus, we only need to focus on keyword privacy. It is the first algorithm known to be suitable for signing as well as encryption, and was one of the first great advances in public key cryptography. RSA is widely used in electronic commerce protocols, and is believed to be secure given sufficiently long keys and the use of up-to-date implementations.

4.3 Ranking and Best File Identification

Ranking the best file is done by calculating the ratio between the frequency and the total number of keywords. The value is calculated and compared with the rest of the values. The maximum valued files are ranked in order. The files are retrieved to the user as the index data of all the files are maintained in the index of the main cloud. The ranking is done on the user side, which may bring in huge computation and post processing overhead. Moreover, sending back all the files consumes large undesirable bandwidth. The best file identification is achieved using Top k query process.

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 inner product of query vector with data vector. However, directly outsourcing data vector or query vector will violate index privacy or search privacy. The maximum ranked values are obtained using term frequency calculation. The files are kept in the ascending order. The best files are given as output to the main cloud server. The main cloud server retrieves top files and given as output to the user.

4.4 Security Analysis

Cloud security architecture is only effective if the correct defensive implementations are in place. Efficient cloud security architecture should recognize the issues that will arise with security management. The security management addresses these issues with security controls. These controls are put in place to safeguard any weaknesses in the system and reduce the effect of an attack.

The cloud server should not learn the plaintext of either the data files or the searched keywords. The new scheme embeds the encrypted relevance scores in the searchable index in addition to file ID. Thus, the encrypted scores are the only additional information that the adversary can utilize against the security guarantee, i.e., keyword privacy and file confidentiality.

Due to the security strength of the file encryption scheme, the file content is clearly well protected. The user gets the ranked results without letting cloud server learn any additional information more than the access pattern and search pattern. 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.

1. 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.
2. Privacy-Preserving: To prevent cloud server from learning additional information from dataset and index, and to meet privacy requirements.
3. Efficiency: Above goals on functionality and privacy should be achieved with low communication and computation overhead.

5. Future Directions

There are possible improvements and undergoing efforts that will appear in the future work. Firstly, the user side of proposed system will be implemented on mobile devices running Android and iOS operating systems since the potential application scenario envisions that users access the data anywhere and anytime. Secondly, the proposed method will be tested on a real dataset in order to compare the performance of our ranking method with the ranking methods used in plain datasets that do not involve any security or privacy-preserving techniques.

6. Conclusion

The problem of solving efficient ranked keyword search is to achieve the effective utilization of remotely stored encrypted data in Cloud Computing. A basic scheme shows that by following the same existing searchable encryption framework, it is very inefficient to achieve ranked search. This appropriately weaken the security guarantee, resort to the newly developed encrypted algorithms, which allows the efficiency in cloud. Investigations of privacy and efficiency guarantees of proposed schemes is given, and experiments on the real-world dataset shows how our proposed schemes introduce low overhead on both computation and

communication. It is a secure and privacy preserving, while correctly realizing the goal of ranked keyword search.

References

[1] Cong Wang, Ning Cao, Jin Li, Kui Ren and Wenjing Lou, Enabling Secure and Efficient Ranked Keyword Search over Outsourced Cloud Data,2012.
[2] S. Kamara and K. Lauter, Cryptographic cloud storage, Workshop on Real-Life Cryptographic Protocols and Standardization 2010, January 2010.
[3] P. Golle, J. Staddon, and B. R. Waters, Secure Conjunctive Keyword Search over Encrypted Data, 2004.
[4] Swaminathan, Y. Mao, G.-M. Su, H. Gou, A. L. Varna, S. He, M. Wu, and D. W. Oard, Confidentiality-preserving rank ordered search, 2007.
[5] Boldyreva, N. Chenette, Y. Lee, and A. O'Neill, Order-preserving symmetric encryption, Springer, 2009.
[6] Y.-C. Chang and M. Mitzenmacher, Privacy preserving keyword searches on remote encrypted data, 2005.
[7] S. Zerr, D. Olmedilla, W. Nejdl, and W. Siberski, Top-k retrieval from a confidential index, 2009.
[8] D. Boneh, G. D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search, Springer, 2004.
[9] Y. H. Hwang and P. J. Lee, Public Key Encryption with Conjunctive Keyword Search and Its Extension to a Multi-User System, 2007.