International Journal of Research in Advent Technology, Vol.5, No.4, April 2017
E-ISSN: 2321-9637
Available online at www.ijrat.org

An Imperial Study on Efficient and Reliable Ranked Keyword search Method

Pooja V. Tundurwar [1], Prof. Sachin A. Murab [2], Prof. M.G.Ingle [3]

1 PG Student, Department of CSE, JCOET Yavatmal, India, pooja.tundurwar07@gmail.com,
2 Professor, Department of CSE, JCOET Yavatmal, India, sachinmurab21@gmail.com,
3 Professor, Department of CSE, JCOET Yavatmal, India., mangeshingle20@gmail.com

Abstract- Cloud data owners prefer to outsource documents in an encrypted form for the purpose of privacy preserving. Therefore it is essential to develop efficient and reliable cipher text search techniques. One challenge is that the relationship between documents will be normally concealed in the process of encryption, which will lead to significant search accuracy performance degradation. Also the volume of data in data centers has experienced a dramatic growth. This will make it even more challenging to design cipher text search schemes on large volume of encrypted data. In this paper, a hierarchical clustering method is proposed to support more search semantics and also to meet the demand for fast cipher text search within a big data environment. The proposed hierarchical approach clusters the documents based on the minimum relevance threshold, and then partitions the resulting clusters into sub-clusters until the constraint on the maximum size of cluster is reached. In the search phase, this approach can reach a linear computational complexity against an exponential size increase of document collection. In order to verify the authenticity of search results, a structure called minimum hash sub-tree is designed in this paper. The results show that with a sharp increase of documents in the dataset the search time of the proposed method increases linearly whereas the search time of the traditional method increases exponentially. Furthermore, the proposed method has an advantage over the traditional method in the rank privacy and relevance of retrieved documents.

Keywords: Cloud computing, ranked search, multi-keyword search, hierarchical clustering, security.

1. INTRODUCTION

Data volume in cloud storage facilities is experiencing a dramatic increase. Although cloud server providers (CSPs) claim that their cloud service is armed with strong security measures, security and privacy are major obstacles preventing the wider acceptance of cloud computing service.

A traditional way to reduce information leakage is data encryption. However, this will make server-side data utilization. In the recent years, researchers have proposed many ciphertext search schemes by incorporating the cryptography techniques. These methods have been proven with provable security, but their methods need massive operations and have high time complexity and the relationship between documents is concealed in the above methods. The relationship between documents represents the properties of the documents and hence maintaining the relationship is vital to fully express a document. Due to the blind encryption, this important property has been concealed in the traditional methods. Therefore, proposing a method which can maintain and utilize this relationship to speed the search phase is desira. Thus, a verifiable mechanism should be provided for users to verify the correctness and completeness of the search results.

2. RELATED WORK

A vector space model is used and every document is represented by a vector, which means every document can be seen as a point in a high dimensional space. Due to the relationship between different documents, all the documents can be divided into several categories. The search time can be largely reduced by selecting the desired category and abandoning the irrelevant categories. Comparing with all the documents in the dataset, the number of documents which user aims at is very small. Due to the small number of the desired documents, a specific

category can be further divided into several sub-categories. Instead of using the traditional sequence search method, a backtracking algorithm is produced to search the target documents. Cloud server will first search the categories and get the minimum desired sub-category. Then the cloud server will select the desired k documents from the minimum desired sub-category. The value of k is previously decided by the user and sent to the cloud server. If current sub-category can not satisfy the k documents, cloud server will trace back to its parent and select the desired documents from its brother categories. This process will be executed recursively until the desired k documents are satisfied or the root is reached. To verify the integrity of the search result, a verifiable structure based on hash function is constructed. Every document will be hashed and the hash result will be used to represent the document. The hashed results of documents will be hashed again with the category information that these documents belong to and the result will be used to represent the current category. Similarly, every category will be represented by the hash result of the combination of current category information and sub-categories information. A virtual root is constructed to represent all the data and categories. The virtual root is denoted by the hash result of the concatenation of all the categories located in the first level. The virtual root will be signed so that it is verifiable. To verify the search result, user only needs to verify the virtual root, instead of verifying every document.

3. PROPOSED SYSTEM

We investigate the problem of maintaining the close relationship between different plain documents over an encrypted domain and propose a clustering method to solve this problem. We proposed the Hierarchical Clustering architecture to speed up server-side searching phase. Accompanying with the exponential growth of document collection, the search time is reduced to a linear time instead of exponential time. We design a search strategy to improve the rank privacy. This search strategy adopts the backtracking algorithm upon the above clustering method. With the growing of the data volume, the advantage of the proposed method in rank privacy tends to be more apparent. By applying the Merkle hash tree and cryptographic signature to authenticated tree structure, we provide a verification mechanism to assure the correctness and completeness of search results.

Advantages:

1. We proposed the Hierarchical Clustering architecture to speed up server-side searching phase. Accompanying with the exponential growth of document collection, the search time is reduced to a linear time instead of exponential time.

2. We design a search strategy to improve the rank privacy. This search strategy adopts the backtracking algorithm upon the above clustering method. With the growing of the data volume, the advantage of the proposev method in rank privacy tends to be more apparent.

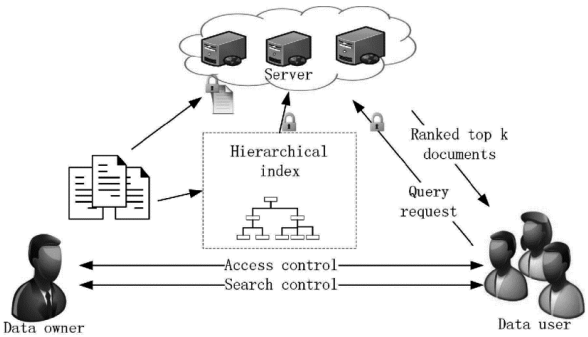

Figure 1: Proposed System Diagram

4. EXISTING SYSTEM

A traditional way to reduce information leakage is data encryption. However, this will make server-side data utilization, such as searching on encrypted data, become a very challenging task. In the recent years, researchers have proposed many cipher text search schemes by incorporating the cryptography techniques. These methods have been proven with provable security, but their methods need massive operations and have high time complexity. Therefore, former methods are not suitable for the big data scenario where data volume is very big and applications require online data processing. In addition, the relationship between documents is concealed in the above methods. The relationship between documents represents the properties of the documents and hence maintaining the relationship is vital to fully express a document.

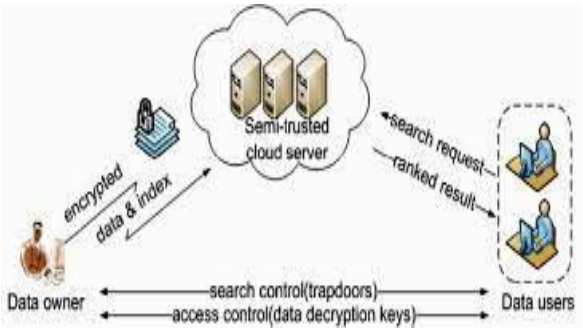

Figure 3: Existing system diagram

Disadvantages:

1. If you are stored data in the cloud after you want any file it takes more time because we are not assign index.

2. If you kept any file in the cloud not provides security.

SYSTEM REQUIREMENTS:

SOFTWARE REQUIREMENTS:

Operating System	- Windows
Language	- J2EE
Database	- MySQL
Database Connectivity	- JDBC

HARDWARE REQUIREMENTS:

Processor	-	Intel Dual Core
Speed	-	2.9 GHz
RAM	-	2 GB (min)
Hard Disk	-	320 GB

5. IMPLEMENTATION DETAILS

The main objective of the proposed system is to improve the efficiency. We propose a hierarchical method in order to get a better clustering result within a large amount of data collection. The size of each cluster is controlled as a trade-off between clustering accuracy and query efficiency. According to the proposed method, the number of clusters and the minimum relevance score increase with the increase of the levels whereas the maximum size of a cluster reduces. Depending on the needs of the grain level, the maximum size of a cluster is set at each level. Every cluster needs to satisfy the constraints. If there is a cluster whose size exceeds the limitation, this cluster will be divided into several sub-clusters.

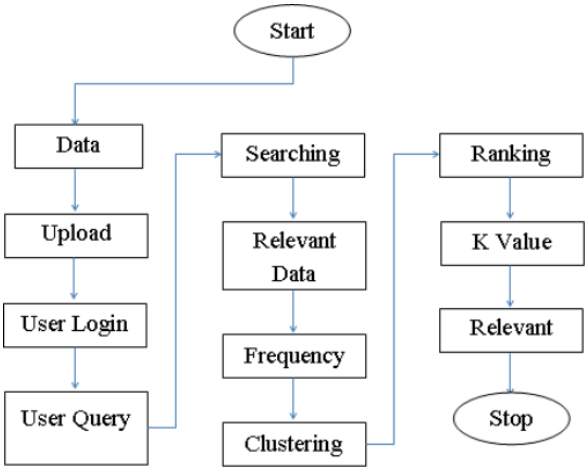

Figure 3: Data Flow Diagram

1. Hierarchical Clustering Architecture:

Hierarchical Clustering architecture is where the data owner builds the encrypted index depending on the dictionary, random numbers and secret key ,the data user submits a query to the cloud server for getting desired documents, and the cloud server returns the target documents to the data user. The vector space model adopted by the Hierarchical Clustering scheme is same as the MRSE while the process of building index is totally different. The hierarchical index structure is introduced into the ins Hierarchical Clustering tead of sequence index. In Hierarchical Clustering every document is indexed by a vector. This architecture mainly consists of following algorithms.

- Keygen(sk,k): It is used to generate the secret key to encrypt index and documents.
- Index(D,sk): Encrypted index is generated in this phase by using the above mentioned secret key. At the same time, clustering process is also included current phase.
- Enc(D,K): The document collection is encrypted by a symmetric encryption algorithm which achieves semantic security.
- Trapdoor(w,sk): It generates encrypted query vector Tw with users input keywords and secret key.
- Search(I,E): In this phase, cloud server compares trapdoor with index to get the top-k retrieval results.

- Dec(Ew,K): The returned encrypted documents are decrypted by the key generated in the first step.

2. Relevance Measure:

In this paper, the concept of coordinate matching]is adopted as a relevance measure. It is used to quantify the relevance of document-query and document-document. It is also used to quantify the relevance of the query and cluster centers. The relevance score between document di and query qw. The relevance score between query qw and cluster center lc i;j. The relevance score between document di and dj.

3. Quality Hierarchical Clustering Algorithm:

We propose a quality hierarchical clustering (QHC) algorithm based on the novel dynamic K-means.As the proposed dynamic K-means algorithm shown in the minimum relevance threshold of the clusters is defined to keep the cluster compact and dense. If the relevance score between a document and its center is smaller than the threshold, a new cluster center is added and all the documents are reassigned. The above procedure will be iterated until k is stable. Comparing with the traditional clustering method, k is dynamically changed during the clustering process.

4. Search Algorithm:

In the search phase, cloud server calculates the relevance score between the query and documents by computing the inner product of the query vector and document vectors and return the target documents to user according to the top k relevance score. The cloud server needs to find the cluster that most matches the query. With the help of cluster index Ic and document classification DC, the cloud server uses an iterative procedure to find the best matched cluster.

MODULE DESCRIPTION

After careful analysis the system has been identified to have the following modules:

1. Data Owner Module.
2. Data User Module.
3. Data Server Module.

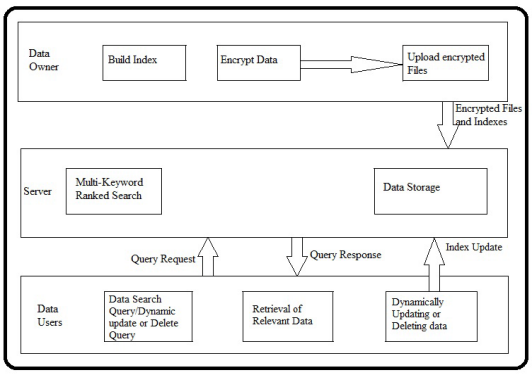

Figure 4: System Architecture

1. Data owner Module:

The data owner is responsible for collecting documents, building document index and outsourcing them in an encrypted format to the cloud server. In this module, the data owners should be able to upload the files [10]. The files are encrypted before the files are uploaded to the cloud. The data owners are provided an option to enter the keywords for the file that are uploaded to the server. These keywords are used for the indexing purpose which helps the search return values very quickly. These files when once available on the cloud, the data users should be able search using the keywords. The data owners will also be provided with a request approval screen so they are able to approve or reject the request that are received by the data users.

2. Data User Module:

The data user needs to get the authorization from the data owner before accessing to the data. Data users are users on this system, who will be able to search files from the cloud that are uploaded by the data owners [11]. Since the files stored on the cloud server could be in huge numbers, there is a search facility provided to the user. The user should be able to do a multi-keyword search on the cloud server.

3. Data Server Module:

The data server provides a huge storage space, and the computation resources needed by ciphertext search. Upon receiving a legal request from the data user, the cloud server searches the encrypted index, and sends back top-k documents that are most likely to match users query. The number k is properly chosen by the data user. Our system aims at protecting data from leaking The data server provides a huge

storage space, and the computation resources needed by ciphertext search. Upon receiving a legal request from the data user, the cloud server searches the encrypted index, and sends back top-k documents that are most likely to match users query. The number k is properly chosen by the data user. Our system aims at protecting data from information to the cloud server while improving the efficiency of ciphertext search. In this model, both the data owner and the data user are trusted, while the cloud server is semi-trusted, which is consistent with the architecture in In other words, the cloud server will strictly follow the predicated order and try to get more information about the data and the index.

6. CONCLUSION

We investigated ciphertext search in the scenario of cloud storage. We explore the problem of maintaining the semantic relationship between different plain documents over the related encrypted documents and give the design method to enhance the performance of the semantic search. We propose the MRSE-HCI architecture to adapt to the requirements of data explosion, online information retrieval and semantic search. At the same time, a verifiable mechanism is also proposed to guarantee correctness and completeness of search results.

REFERENCES

1. D. X. D. Song, D. Wagner, and A. Perrig, ”Practical techniques for searches on encrypted data,” in Proc. S & P, BERKELEY,CA, 2000, pp. 44-55.
2. D. Boneh, G. Di Crescenzo, R. Ostrovsky, and G. Persiano,”Public key encryption with keyword search,” in Proc. EUROCRYPT, Interlaken, SWITZERLAND, 2004, pp. 506-522.
3. Y. C. Chang, and M. Mitzenmacher, ”Privacy preserving keyword searches on remote encrypted data,” in Proc. ACNS, ColumbiaUniv, New York, NY, 2005, pp. 442-455.
4. R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky, ”Searchable symmetric encryption: improved definitions and efficient constructions,” in Proc. ACM CCS, Alexandria, Virginia, USA,2006, pp. 79-88.
5. M. Bellare, A. Boldyreva, and A. O’Neill, ”Deterministic and efficiently searchable encryption,” in Proc. CRYPTO, Santa Barbara, CA, 2007, pp. 535-552.
6. D. Boneh, and B. Waters, ”Conjunctive, subset, and range queries on encrypted data,” in Proc. TCC, Amsterdam, NETHERLANDS, 2007, pp. 535-554.
7. D. X. D. Song, D. Wagner, and A. Perrig, ”Practical techniquesfor searches on encrypted data,” in Proc. S & P 2000, BERKELEY, CA, 2000, pp. 44-55.
8. E.-J. Goh, Secure Indexes, IACR Cryptology ePrint Archive, vol.2003, pp. 216. 2003.
9. C. Wang, N. Cao, K. Ren, and W. J. Lou, Enabling Secure and Efficient Ranked Keyword Search over Outsourced Cloud Data, IEEE Trans. Parallel Distrib. Syst., vol. 23, no. 8, pp. 1467-1479,Aug. 2012.
10. Iternational Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 02 Issue: 03 | June-2015, www.irjet.net .
11. In International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Volume 4 Issue 11, November 2015 www.ijsr.net Licensed Under Creative Commons Attribution CC BY “A Survey on Multi-Keyword Ranked Query Search over Encrypted Cloud Storage.”.