ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

A Survey on Privacy-Preserving Ranked Keyword Search Method

Harshal Khona, Akash Gulhane, Rishab Sethi, Dinesh Yadav, Shalaka Deore

B.E Student, Department of CSE, M.E.S College of Engg, Savitribai Phule Pune University, Pune, India

B.E Student, Department of CSE, M.E.S College of Engg, Savitribai Phule Pune University, Pune, India

B.E Student, Department of CSE, M.E.S College of Engg, Savitribai Phule Pune University, Pune, India

B.E Student, Department of CSE, M.E.S College of Engg, Savitribai Phule Pune University, Pune, India

Assistant Professor, Department of CSE, M.E.S College of Engg, Savitribai Phule Pune University, Pune, India

ABSTRACT: Cloud data users prefer to outsource documents in an encrypted form for the purpose of preserving privacy. Therefore it is extremely important to develop efficient and reliable cipher-text search techniques. A major challenge is that the relationship between documents will be concealed in the process of encryption, which will lead to a significant degradation in search accuracy performance. In this paper, a hierarchical clustering method is proposed to support more search semantics and to meet the need for fast cipher-text search within a big data environment. The proposed hierarchical approach forms clusters of the documents based on a minimum relevance threshold, and later partitions the resulting clusters into sub-clusters until the constraint on the max size of cluster is reached. In the search phase, this approach does fairly good by reaching a linear computational complexity against an exponential size growth in the number of documents. In order to verify the authenticity and correctness of search results, minimum hash sub-tree is designed.

KEYWORDS: Cipher Text Search; Cloud Computing;Hierarchical Clustering; Multi-Keyword Search; Ranked Search;Security.

I. INTRODUCTION

In the era of big data, huge amount of data is produced world-wide. Enterprises choose to outsource their large amount of data to cloud facility in order to reduce the cost of data management and storage facility spending. As a result of this, data volume in cloud storage facilities is experiencing a dramatic increase. Although cloud server providers claim that their cloud service is armed with strong security measures, security and privacy are major obstacles preventing the broad acceptance of cloud computing service.

A traditional approach of reducing leakage in information is data encryption. However, this makes the server-side data utilization, such as searching on encrypted data, a very problematic task. In recent years, researchers have proposed many cipher-text search schemes by incorporating the techniques of cryptography. These methods have been proven with good security, but their methods need massive operations to be performed and also have high time complexity. Therefore, former methods are not suitable for the big data scenario where data volume is huge and applications require online processing of data. In addition, there is concealment in the relationship between documents in the above methods. The relationship between documents represents the properties of the documents and hence maintaining this relationship is necessary to fully express a document. For example, the relationship can be used to express its class. If a document is independent of any other documents except those documents that are related to business, then it is easy for us to assert this document belongs to the category of the business. Due to the blind encryption, this vital property has been concealed in the traditional methods. Therefore, proposing a method which can utilize and maintain this relationship to speed the search phase is desirable.

Also, due to failure of software/hardware, and storage corruption, data search results may contain damaged data or may have been distorted by intruder. Therefore, a mechanism should be provided for users to verify the correctness as well as the completeness of search results.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

In this paper, every document is represented by a vector by using a vector space model, which means every document can be seen as a small point in a space. All the documents can be divided into several categories because of the relationship between different documents. In other words, the points with short distance in the high dimensional space can be classified into a specific category. The search time can be highly reduced by selecting the desired category and rejecting the irrelevant categories. The number of documents which user aims at is very small compared to the number of documents in the dataset. Due to the limited number of the desired documents, a specific category can be further divided into different sub-categories. Instead of using the traditional search methods, a backtracking algorithm is produced to search the targeted documents. Cloud server first searches the categories and gets the minimum desired sub-category. Then the cloud server selects the desired k documents from the minimum desired sub-category. The user decides the value of k and sends it to the cloud server. If current sub-category cannot satisfy the k documents, cloud server traces back to its parent and selects the desired documents from its brother categories. This process executes recursively until the desired k number of documents are either satisfied or the root node is reached. To verify the integrity of the search result, hash function is used. All documents will be hashed and the hash result will be used to represent the document. The results of documents will be hashed again with the information of category that these documents belong to and the result will represent the current category. Similarly, every category can be represented by the hash result of the current category information as well as the sub-categories information.
A virtual root is constructed to represent 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. This virtual root will be signed so that it is verifiable. To verify the results of search, user now only needs to verify the virtual root, instead of verifying all the documents.

II. RELATED WORK

Single Keyword Searchable Encryption

The notion of searchable encryption was first introduced by Song. The proposal was to encrypt the words in the document independently. This has a high searching cost due to the word by word scanning of the whole data. Cash et al. recently designed and implemented an efficient data structure. Due to the lack of rank mechanism, users require a lot of time to select the document when large number of documents contain the query keyword. Wang et al. used encrypted invert index to achieve secure ranked keyword search over the documents which were encrypted. In the search phase, the cloud server calculates the relevance score between documents and the query. In this way, related documents are ranked according to their score (relevance) and users can get topk relevant results. Boneh et al designed a searchable encryption construction, first of its type, where anyone can use public key to write to the data stored on server but private key is provided only to the authorized users and only these users can search. However, thesemethodsmentioned above only support single keyword search.

Multiple Keyword Searchable Encryption

To enhance search predicates, different conjunctive keyword search methods have been proposed. These methods have a large overhead. Pang et al. proposed asecure search technique based on vector space model. The efficiency and security of this technique is inefficient due to the lack of the security analysis for practical search performance. Cao et al. presented a novel method to solve the issue of multi-keyword ranked search over encrypted cloud data. But the drawback being that the search time of this technique grows exponentially accompanying with the exponential increasein the size of the document collections. Sun et al. gave a new architecture which achieves better search efficiency. However, the relevance between documents is ignored. As a result, expectations of the user cannot be fulfilled well. For example: given a query containing Cell and Phone, only the documents containing both these keywords will be retrieved by traditional methods. But by taking the semantic relationship between the documents into consideration, the documents containing Mobile and Phone should also be retrieved. As a result, the second result is better at meeting the expectations of the user.

III. PROPOSED WORK

In this paper, we are proposing a multi-keyword ranked search over the encrypted data based on hierarchical clustering index (MRSE-HCI) to maintain a close relationship between various plain documents in order to enhance the search efficiency over the encrypted domain. In this proposed architecture, the search time grows linearly accompanying with an exponential growing size of data collection. This idea is derived from the observation that user's

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

retrieval is usually concentrated on a specific field. This was, we can speed up the process of searching results by computing the relevance score between the query and documents belonging to the same specific field as that of the query. As a result, only those documents which are classified to the field specified by users query will be evaluated to get their relevance score. As the irrelevant fields are ignored, the search speed is enhanced.

We look into the problem of maintaining the close relationship between various documents over an encrypted domain and also propose a clustering method to solve the problem. According to our proposed clustering method, every document will be classified dynamically into a specific cluster based on a constraint on the relevance score (minimum) between different documents. The relevance score is used to evaluate the relationship between different documents in the dataset. Due to the addition of new documents to a cluster, the constraint on the cluster may or may not be broken. If one of the newly added documents breaks the constraint, a new cluster will be added and the current document will be chosen as a temporary value of this cluster center. Then all the documents added so far will be reassigned and all the cluster centers will be re-elected. Therefore, the number of documents in the dataset and the close relationship between different plain documents is directly dependent on the number of clusters. In other words, the cluster centers are created dynamically as new clusters can be formed after the addition of new documents and the number of clusters is decided by the dataset.

We propose a hierarchical method to get a better clustering result within the big data environment. The cluster-size is controlled as a trade-off between query efficiency and clustering accuracy. According to the proposed method, the number of clusters and the relevance score increase with the increase in the number of levels whereas there is a reduce in the size of the clusters. Depending on the needs of the grain level, at each level, a maximum size of the cluster is set. All the clusters need to satisfy the constraint. If a cluster exceeds the limitation of maximum size, then this cluster is divided into several sub-clusters.

We implement a search strategy to improve the rank privacy. In the search phase, the cloud server first computes the relevance score between the search query and first level cluster centers and then chooses the nearest cluster. This process is iterated until the smallest cluster has been found. The cloud server then computes the relevance score between search query and documents present in the smallest cluster. If the smallest cluster has fewer documents than the number of desired documents which was previously decided by the user, the cloud server searches the brother clusters of the smallest cluster by backtracking to the parent cluster. This process is iterated until the number of desired documents i.e. the value of k is satisfied or the root node is reached. The rank privacy is enhanced due to the special search procedures as the rankings of documents among their search results are different from the rankings derived from traditional sequence search.

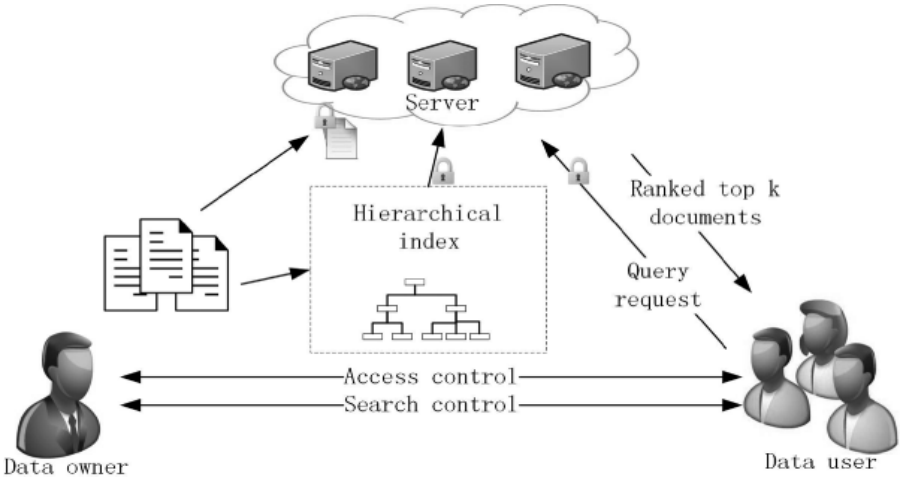
For further improvement, we also construct a verifiable tree structure upon the hierarchical clustering method to verify the integrity of the search result. This authenticated tree structure mainly takes the advantage of the Merkle hash tree and cryptographic signature. Every document will be hashed and the hash result will be used as the representative of the document. The smallest cluster is represented by the hash value of the combination of concatenation of documents included in the smallest cluster and own category information. The parent cluster is represented by the hash result of the combination of the concatenation of its children and own category information. A virtual root is added and represented by the hash value of the concatenation of the categories located in the first level. Also, the virtual root will be signed so that user can achieve the goal of verifying the search result by verifying the virtual root.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016



IV. ARCHITECTURE

In this section, we introduce the *MRSE-HCI* scheme. The vector space model adopted by the *MRSE-HCI* scheme is same as the *MRSE*, while the process of building index is totally different. The hierarchical index structure is used instead of the sequence index into the *MRSE-HCI*. In *MRSE-HCI*, all documents are indexed by a vector. All dimensions of the vector stand for a keyword and the value represents whether the keyword appears in the document or not. Similarly, the query is also represented by a vector. 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 relevance score.

Due to the fact that all the documents outsourced to the cloud server are encrypted, the semantic relationship is lost between plain documents over the encrypted documents. In order to maintain this relationship between plain documents over the encrypted documents, a clustering method is used by clustering the related index vectors of documents. Every document vector is viewed as a point in the high n-dimensional space. With the length of vectors being normalized, we know that the distance of cluster points in the n-dimensional space reflect the relevance of corresponding documents. In other words, points of high relevant documents are very close to each other in the high n-dimensional space. As a result, we can cluster the documents based on the distance measure.

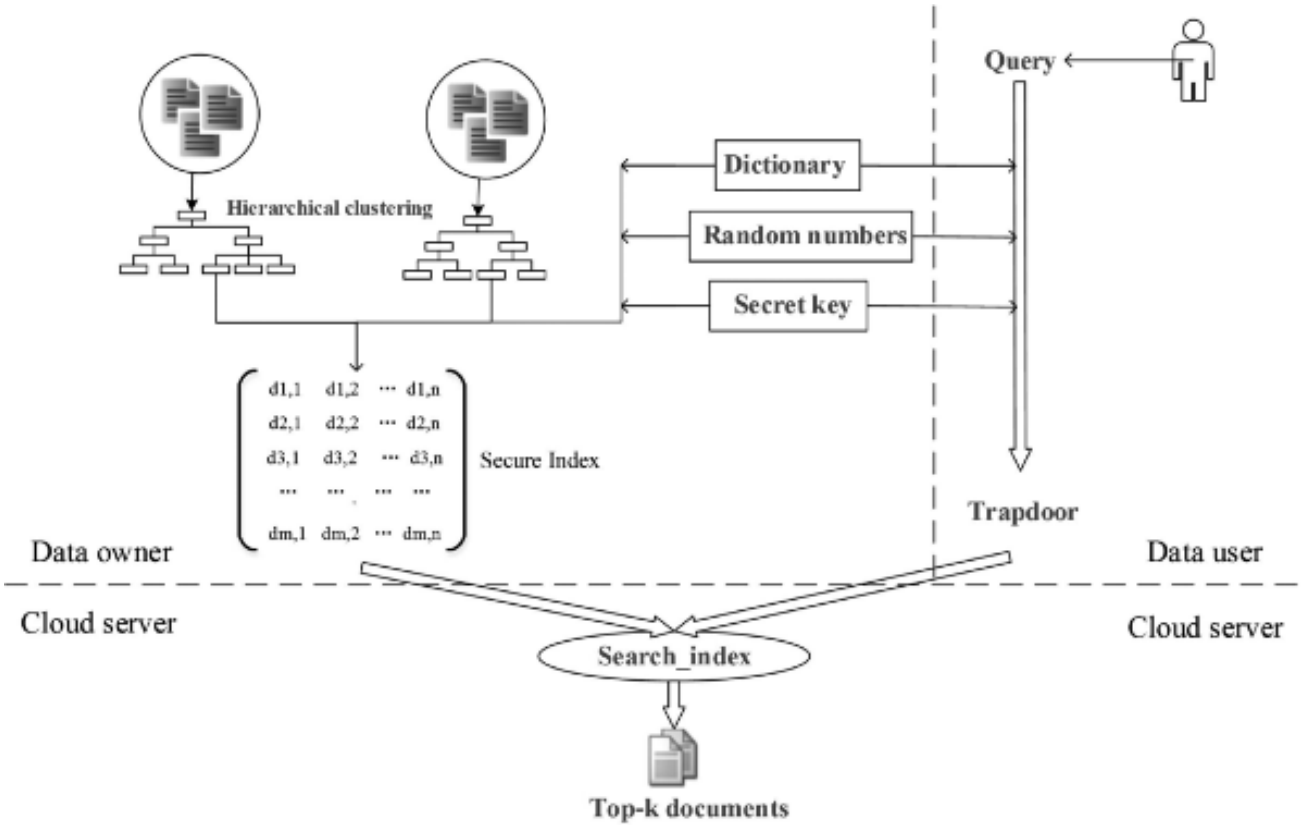
As the volume of data in the data centers have experienced a dramatic growth, conventional sequence search approach will be very inefficient. To promote the search efficiency, a hierarchical clustering method is proposed. The proposed method clusters the documents based on the minimum relevance threshold at different stages, and then partitions the resulting clusters into sub-clusters until the constraints on the max size of cluster are reached. After receiving a legal request, cloud server will only search the related indexes layer by layer instead of scanning all indexes.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016



V. CONCLUSION AND FUTURE WORK

In this paper, we investigated cipher-text search in cloud storage scenario. We explored the problem of maintaining the relationship between different documents over the related encrypted documents and enhance the performance of the semantic search. We also proposed the *MRSE-HCI* architecture to adapt to the requirements of online information retrieval and semantic search. Also, a verifiable mechanism is proposed to guarantee the completeness as well as the correctness of search results. In addition, we analyse the search efficiency and security under popular threat models. An experimental platform is used to evaluate the search efficiency, accuracy, and rank security. This experiment result proves that the proposed architecture solves the multi-keyword ranked search problem as well as brings an improvement in rank security, search efficiency, and the relevance between documents.

REFERENCES

1. Chi Chen, Xiaojie Zhu, Peisong Shen "An Efficient Privacy-Preserving Ranked Keyword Search Method", IEEE Transactions on Parallel and Distributed Systems, Vol. 27, No. 4, April 2016
2. H. Pang, J. Shen and R. Krishnan 'Privacy-preserving similarity-based text retrieval' , *ACM Trans. Internet Technol., vol. 10, no. 1, p. 39, Feb., 2010*
3. C. Martel, G. Nuckolls, P. Devanbu, M. Gertz, A. Kwong and S. G. Stubblebine 'A general model for authenticated data structures' *Algorithmica, vol. 39, no. 1, pp. 21-41, May, 2004*
4. M. Naor and K. Nissim 'Certificate revocation and certificate update'*IEEE J. Sel. Areas Commun, vol. 18, no. 4, pp. 561-570, Apr., 2000*
5. H. Pang and K. Mouratidis "Authenticating the query results of text search engines"*Proc. VLDB Endow., vol. 1, no. 1, pp. 126-137, Aug., 2008*
6. Z. X. Huang "Extensions to the k-means algorithm for clustering large data sets with categorical values"*Data Min. Knowl. Discov., vol. 2, no. 3, pp. 283-304, Sep., 1998*
7. R. X. Li, Z. Y. Xu, W. S. Kang, K. C. Yow and C. Z. Xu "Efficient Multi-keyword ranked query over encrypted data in cloud computing"*Futur. Gener. Comp. Syst., vol. 30, pp. 179-190, Jan., 2014*
8. G. Craig "Fully homomorphic encryption using ideal lattices"*Proc. 41st Annu. ACM Symp. Theory Comput., vol. 9, pp. 169-178, 2009*
9. S. Jarecki, C. Jutla, H. Krawczyk, M. Rosu and M. Steiner "Outsourced symmetric private information retrieval"*Proc. ACM SIGSAC Conf. Comput. Commun. Secur., pp. 875-888, Nov., 2013*

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2016

10. M. Chase and S. Kamara "Structured encryption and controlled disclosure"*Proc. Adv. Cryptol., pp. 577-594, 2010*
11. R. Curtmola, J. Garay, S. Kamara and R. Ostrovsky "Searchable symmetric encryption: Improved definitions and efficient construc-tions"*Proc. 13th ACM Conf. Comput. Commun. Secur., pp. 79-88, 2006*
12. S. Kamara, C. Papamanthou and T. Roeder "Dynamic searchable symmetric encryption"*Proc. Conf. Comput. Commun. Secur., pp. 965-976, 2012*
13. D. Cash, S. Jarecki, C. Jutla, H. Krawczyk, M. Rosu and M. Steiner "Highly-scalable searchable symmetric encryption with support for Boolean que-ries"*Proc. Adv. Cryptol,., pp. 353-373, 2013*