RESEARCH ARTICLE OPEN ACCESS

Quality Hierarchical Cluster Algorithm to Verify Search Patterns in Cloud Data

*Dr. P. Julia Grace[1],T.P.Padmini[2]

[1]*AssistantProfessor & Research Supervisor, Department of Computer Science, JBAS College for Women, Chennai.*

[2] *M.Phil., Computer Science Scholar,Mother Teresa women's University, kodaikanal.*

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. This paper challenges 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 that can provide efficient and reliable online information retrieval 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. Experiments have been conducted using the collection set available. 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.The Cloud storage is illustrated as a working model in school for Payroll management and Appraisal form.

Keywords: *cipher text search schemes, cloud data set, hierarchal approach, rank privacy.*

Date of Submission: 02-08-2017 Date of acceptance: 14-08-2017

I. INTRODUCTION

As cloud computing has become prevalent, sensitive information are being centralized in the cloud server. To have privacy and security of data in the cloud server the data is stored in encrypted form. 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. Therefore it is important to develop an efficient and reliable cipher text search techniques. Although traditional searchable encryption schemes allow a user to securely search over encrypted data, these methods need massive operations, have high time complexity, very costly, and not fast enough for big data environment.

To overcome the above problems, hierarchical cluster method has been proposed. This proposed method supports more search semantics and also to meet the demand for fast cipher text search within a big data environment. The result in the proposed method shows that with a sharp increase of documents in the dataset the search time increased linearly whereas the search time in the traditional method increases exponentially. In this thesis we have also included methods to verify the authenticity of search results by a structure called minimum hash sub-tree. Through rigorous analysis, we show that the proposed method provides a verification mechanism to assure the correctness and completeness of search results. It also has an advantage over the traditional method in the rank privacy and relevance of retrieved documents. The above idea has been implemented as working model in SBOA School and Junior College, Chennai It enables self-appraisal for the staff and to maintain confidential report about the staff. This application will help the management to promote the staff based on their appraisal innovatively. It also helps to generate pay slip for the staff confidentially. It is possible only when the staff provides authenticated information approved by the Administrator.

II. DESIGN GOALS

Enterprises and users whoown a large amount of data usually choose to outsource their precious data to cloud facility in order to reduce datamanagement cost and storage facility spending. As a result,data volume in cloud storage facilities is experiencing adramatic increase. Although cloud server providers (CSPs)claim that their cloud service is armed with strong securitymeasures, security and privacy are major obstacles preventingthe wider acceptance of cloud computing service.Due to software/hardware failure, and storage corruption, data search results returning to the users may contain damaged data or have been distorted by the malicious administrator or intruder. Thus, a verifiable mechanism should be provided for users to verify the correctness and completeness of the search results.

In short, our contributions can be summarized as follows:

1) We investigate the problem of maintaining the closerelationship between different plain documents overan encrypted domain and propose a clusteringmethod to solve this problem.

2) We proposed the MRSE-HCI architecture to speedup server-side searching phase. Accompanying withthe exponential growth of document collection, thesearch time is reduced to a linear time instead ofexponential time.

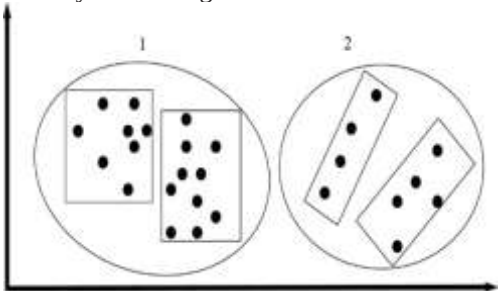
3) We design a search strategy to improve the rank privacy.This search strategy adopts the backtrackingalgorithm upon the above clustering method. Withthe growing of the data volume, the advantage ofthe proposed method in rank privacy tends to bemore apparent.

4) By applying the Merkle hash tree and cryptographicsignature to authenticated tree structure, we providea verification mechanism to assure the correctnessand completeness of search resultsreelected. Therefore, the number of clusters depends on thenumber of documents in the dataset and the close relationshipbetween different plain documents. In other words, thecluster centers are created dynamically and the number ofclusters is decided by the property of the dataset.We propose a hierarchical method in order to get a betterclustering result within a large amount of data collection.The size of each cluster is controlled as a trade-off betweenclustering accuracy and query efficiency. According to theproposed method, the number of clusters and the minimumrelevance score increase with the increase of the levelswhereas themaximum size of a cluster reduces. Dependingon the needs of the grain level, the maximum size of a clusteris 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.

III. METHODOLOGY

3.1 Quality Hierarchical Clustering Algorithm

A lot of hierarchical clustering methods has been proposed. However all of these methods are not comparable to the partition clustering method in terms of time complexity performance. In partition clustering algorithms, the k is fixed, which cannot be applied to the situation of dynamic number of cluster centers. A quality hierarchical clustering (QHC) algorithm based on the novel dynamic K-means is proposed. 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. This is why it iscalled dynamic K-means algorithm. The QHC algorithm is illustrated well with an example. Every cluster will be checked on whether its size exceeds the maximum number TH or not. If the answer is "yes", this "big" cluster will be split into child clusters which are formed by using the dynamic K-means on the documents of this cluster. This procedure will be iterated until all clusters meet the requirement of maximum cluster size. Clustering procedure is illustrated in All the documents are denoted as points in a coordinate system. These points are initially partitioned into two clusters by using dynamic K-means algorithm when the k ¼ 2. These two bigger clusters are depicted by the elliptical shape. Then these two clusters are checked to see whether their points satisfy the distance constraint. The second cluster does not meet this requirement, thus a new cluster center is added with k ¼ 3 and the dynamic K-means algorithm runs again to partition the second cluster into two parts. Then the data owner checks whether these clusters size exceed the maximum number TH. Cluster 1 is split into two sub-clusters again due to its big size. Finally all points are clustered into four clusters as depicted by the rectangle.



Clustering process

3.2 Search Algorithm

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. Following instance demonstrates how to get matched one:

1) The cloud server computes the relevance score between Query Tw and encrypted vectors of the first level cluster centers in cluster index (Ic), then choosesthe i^{th} cluster center Ic;1;i which has the highest score.

2) The cloud server gets the child cluster centers of the cluster center, then computes the relevance score between Tw and every encrypted vectors of child cluster centers, and finally gets the cluster center Ic;2;I with the highest score. This procedure will be iterateduntil that theultimate cluster center Ic;l;j in lastlevel l is achieved.In the situation depicted bythere are nine documentswhich are grouped into three clusters. After calculatingthe relevance score withtrapdoor (Tw) , cluster 1, which isshown within the box of dummy line is found to bethe best match. Documents d1,d3 ,d9 belong to cluster 1, thentheir encrypted document vectors in the Id are extracted outto compute the relevance score with Tw. The figure 3.2.1 shows retrieval process and figure 3.2.2 shows the building minimum hash sub tree algorithm

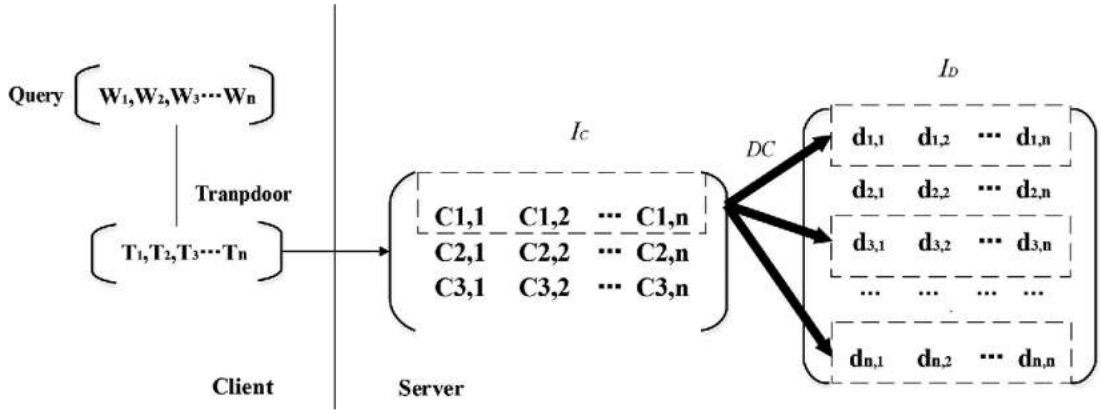


Fig 3.2.1 Retrieval Process of documents in clusters used in our example

Algorithm *Building-minimum hash sub-tree(MHST)*
1 build hash tree based on hierarchical clustering result
2 **for every** leaf node *i*,
3 calculate its hash value:
4 **while** not tree root
5 **for** every non-leaf node *j*
6 calculate its hash value:
7 construct node $(id_j,)$
7 go to the upper level
8 calculate tree root's hash value:
9 calculate the signature of hash value

Fig 3.2.2 Building-minimum hash sub-tree algorithm.

3.3 Search Result Verification

The retrieved data have high possibility to be wrong since the network is unstable and the data may be damaged due to the hardware/software failure or malicious administrator or intruder. Verifying the authenticity of search results is emerging as a critical issue in the cloud environment. A signed hash tree to verify the correctness and freshness of the search results is designed. The data owner builds the hash tree based on the hierarchical index structure. The figure 3.3.1 shows the processing minimum hash sub tree algorithm and figure 3.3.2 shows the authentication for hierarchical clustering index.

Algorithm *Processing-minimum hash sub-tree(MHST)*

```
1 for every leaf node j in the matched cluster
2   add its hash value to the MHST
3 change j to its father node in the upper level
4 while not tree root
5   add j's hash value to the MHST
6   for every j's brother node:
7         add hash value of  to the MHST
8   change j to its father node in the upper level
9 add root signature to the MHST
```

Fig 3.3.1Processing-minimum hash sub-treealgorithm.

.

To deal with this impact on the hash tree, a lazy update strategy is designed. For the insertion operation, the corresponding hash value will be calculated and marked as a raw node, while the original nodes in the hash tree will be kept unchanged because the original hash tree still supports document verification except the new document. Only when the new added document is accessed, the hash tree will be updated. Similar concept is used in the deletion operation. The only difference is that the deletion operation will not bring the hash tree update Efficiency and Security.

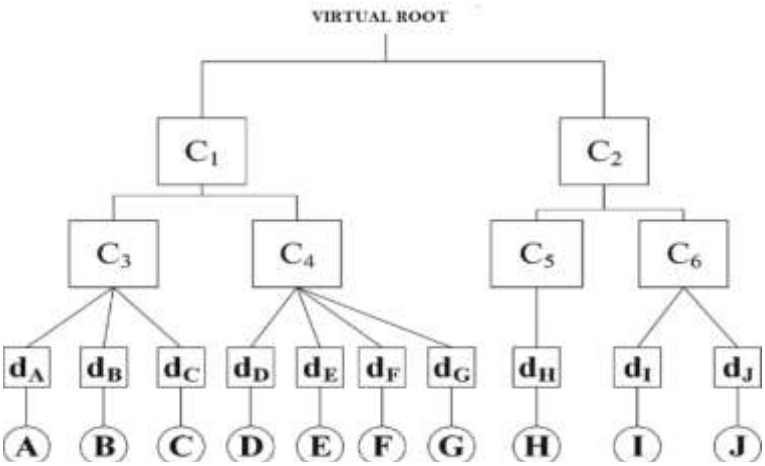


Fig. 3.3.2 Authentication for hierarchical clustering index.

3.4Performance Analysis:

In order to test the performance of MRSE-HCI on real dataset, an experimental platform to test the search efficiency, accuracy and rank privacy is built. Figure 3.4.1 shows Rank privacy for retrieved documents.

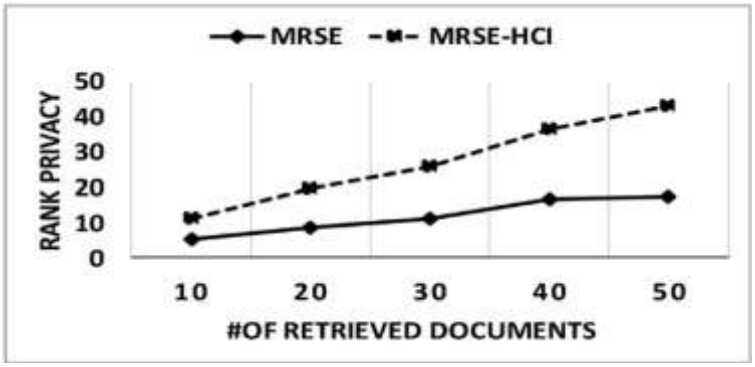


Fig 3.4.1 Rank Privacy

Figure 3.4.1 describes search accuracy by utilizing plaintext search as a standard. From the Figure, we can see that the relevance between query and retrieved documents in MRSE-HCI is slightly lower than that in MRSE. Especially, this gap narrows when the data size increases since a big document data set has a clear category distribution which improves the relevance between query and documents. Figure shows the rank accuracy according to Equation The tradeoff parameter is set to 1, which means there is no bias towards relevance of documents or relevance between documents and query. From the result, it is concluded that MRSE-HCI is better than MRSE in rank accuracy. Figure describes the rank privacy according to Equation In this test, no matter the number of retrieved documents, MRSE _ HCI has better rank privacy than MRSE. This mainly caused by the relevance of documents introduced into search strategy.

In this paper, cipher text search in the scenario of cloud storage is illustrated and 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 is explored. The MRSE-HCI architecture to adapt to the requirements of data explosion, online information retrieval and semantic search is proposed. At the same time, a verifiable mechanism is also proposed to guarantee the correctness and completeness of search results. In addition, Also the search efficiency and security under two popular threat models is analyzed. An experimental platform is built to evaluate the search efficiency, accuracy, and rank security. The experiment result proves that the proposed architecture not only properly solves the multi-keyword ranked search problem, but also brings an improvement in search efficiency, rank security, and the relevance between retrieved documents.

IV. FINAL FINDINGS AND CONCLUSION

The cipher text search in the scenario of cloud storage is illustrated as a working model in SBOA School and Junior College, Chennai. Payroll management should have report generation module for giving salary slips, Appraisal form, feedback form and preparation of Performance reports or Chart. At present this product is run locally and cannot be accessed by everyone. It can be deployed on a web server so that anyone connected to the Internet can access it.

REFERENCES:

[1]. S. Grzonkowski, P. M. Corcoran, T. Coughlin, "Security analysis of authentication protocols for next-generation mobile and CE cloud services", *Proc. IEEE Int. Conf. Consumer Electron. 2011*, pp. 83-87, 2011.

[2]. D. X. D. Song, D. Wagner, A. Perrig, "Practical techniques for searches on encrypted data", *Proc. IEEE Symp. Security Priv.*, pp. 44-55, 2000.

[3]. . Boneh, G. Di Crescenzo, R. Ostrovsky, G. Persiano, "Public key encryption with keyword search", *Proc. EUROCRYPT*, pp. 506-522, 2004.

International Journal of Engineering Research and Applications (IJERA) is **UGC approved** Journal with Sl. No. 4525, Journal no. 47088. Indexed in Cross Ref, Index Copernicus (ICV 80.82), NASA, Ads, Researcher Id Thomson Reuters, DOAJ.

Dr. P. Julia Grace. "Quality Hierarchical Cluster Algorithm to Verify Search Patterns in Cloud Data ." International Journal of Engineering Research and Applications (IJERA), vol. 7, no. 8, 2017, pp. 13–17.