International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 07 | July -2017 www.irjet.net p-ISSN: 2395-0072

Efficient Privacy Preserving Clustering Based Multi Keyword Search

Shital Agrawal [1], Prof. Bharati Dixit[2]

[1,2]*Department of Information Engineering, MIT College of Engineering.*

Abstract - *Recently, cloud computing become more popular, data owners shows their interest or prefer the cloud for storing their personal data, by storing data into the cloud save the economical funds and gives the great flexibility. As providing security to the data is the challenging task for the researchers, for protecting the data sensitive data have to be encoded before outsourcing the data. To accomplish the privacy preserving effective search over encoded cloud information this paper presents the hierarchical clustering approach. Alongside this precise and proficient search over outsourced information, security is additionally considered with user revocation approach. Another more imperative component of this framework is data duplication checking with SHA1 hashing strategy. This strategy will not permit the data owner to store the duplicate data at cloud server. By using the data duplication technique memory overload is reduce on cloud. In this framework for clustering process EM clustering is used, and finally compare the EM clustering algorithm with K-means clustering algorithm. Experimental results demonstrate that the framework has numerous advantages like efficient and effective memory and time utilization, efficient and secure search over encrypted data, secure data storage and data duplication checking.*

Key Words: **Cloud computing, cipher text search, ranked search, multi-keyword search, hierarchical clustering, big data, security.**

1. INTRODUCTION

Now a days cloud computing becomes more interesting methodology in different applications like academics and the industries. Cloud computing has a number of features, for example resource management, economical cost and simple and quick deployment. Due to the tremendous economic advantages cloud computing, most of the organization deployed their cloud centers. Such cloud centers are Elastic Compute Cloud of Amazon, the App Engine of Google, the Azure of Microsoft, and Blue Cloud of IBM. Despite the fact that data is sometime sensitive information, for example, personal records, financial records pass record and so on., because once the data is outsourced, owner can not directly control the data. It is accessible online at some point. The Cloud Service provider (CSP) can keep up the security of such sensitive data by utilizing a few methods like fire ware, virtualization, and Intrusion Detection System (IDS). For this situation, CSP increases full control over such information. But there is no guarantee or full trust on the representative of CSP [1][10]. They can leak or alter the information. That is they can uncover the sensitive data of data owners. So to defeat the issue of security of outsource information, information encryption is one of the best solutions.

The encryption system may have reinforced the data security of cloud data; in any case it has additionally corrupted the proficiency of the data on the grounds that the encryption will decreases the search capacity of the data. Especially in the cloud computing environment, it is illogical for the customer to download and decrypt the entire encoded data from the remote cloud server before a search happens. Along these lines, an proficient plan that backings search over encrypted information in cloud computing turns out to be extremely noteworthy before much organization can exploit the cloud storage [8][9].

Consistently, accessible encryption strategies have been made to give the capacity to explicitly recovering the encrypted reports through a keyword search. Typically, these systems manufacture a protected index structure and outsource it nearby the encoded records to the remote server. Approved customers show their requests as secret trapdoors that are coordinated legitimately with the stored index data. The server uses the got trapdoor to discover the put away file, and gets the coordinating encoded files [10].

Numerous analysts have created many cipher content search framework by utilizing the cryptography strategies. These procedures have demonstrated provable security, however as they requests enormous operations and huge complexity in time. Henceforth already composed systems are not valuable for huge data where size of the data is huge likewise it needs online data processing. Because of the visually impaired encryption, important property has been covered in the customary techniques. Along these lines, proposing a technique which can keep up and use this relationship to speed the search phase is desirable

Proposed system includes following points:

- Propose a clustering method to solve problem of maintaining the close relationship between different plain documents over an encrypted domain.
- Proposed the MRSE-HCI architecture to speed up server side searching phase.
- Design a search strategy to improve the rank privacy.

- Provide a verification mechanism to assure the correctness and completeness of search results.
- Generate ranked top-k documents.
- Provide user revocation method for security.

To achieve above features and for better performance of system, we have used following algorithms and techniques in system.

- AES encryption algorithm and User revocation method for Security
- EM clustering algorithm to enhance the clusters generation accuracy.
- Hierarchical algorithm to generate clusters hierarchically.
- SHA-1 algorithm for hashing
- Perform the data duplication operation for checking the duplicate data.

In this paper study about the related work done, in section II, the proposed approach modules description, mathematical modeling, algorithm and expected result in section III. And at final provide a conclusion in section IV.

2. LITERATURE REVIEW

Chen, Chi, et al.[1], a hierarchical clustering strategy is proposed to support more search semantics furthermore 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 stage, this methodology can achieve a linear computational complexity against an exponential size increment of document collection. All together to confirm the authenticity of search results, a structure called minimum hash sub-tree is designed.

Various well-known [2] authentication protocols are considered with context of next generation mobile and CE network services. The potential weaknesses of current protocol scan be overcome utilizing Zero Knowledge Proof (ZKP) strategies to ensure client passwords so an alternative ZKP protocol, Se Di Ci 2.0, is depicted. This offers mutual and also two-factor authentication that is viewed as more secure against different phishing endeavors than existing trusted outsider protocols. The suitability of such a ZKP protocol for different CE-based cloud computing applications is illustrated.

Wang et al. [3], characterize and solve the issue of secure ranked keyword search over encrypted cloud data. Ranked search enormously upgrades framework usability by enabling search result relevance ranking instead of sending un differentiated results and further guarantees the file retrieval accuracy. In particular, they investigate the statistical measure approach, i.e., relevance score, from data recovery to manufacture secure searchable index, and build up a one-to-many order-preserving mapping strategy to appropriately secure those sensitive score data. The resulting design is able to facilitate efficient server-side ranking without losing keyword privacy.

Pang et al. [4], present a privacy-preserving, similarity-based text retrieval scheme that prevents the server from precisely reproducing the term composition of queries and documents, and anonymize the search results from unauthorized observers. In the meantime, their plan preserves the relevance-ranking of the search server, and empowers accounting of the number of documents that every client opens. The effectiveness of the scheme is verified empirically with two real text corpora.

Cao et al. [5], interestingly, authors characterize and tackle the challenging issue of privacy-preserving multi-keyword ranked search over encrypted information in cloud computing (MRSE). They build up a set of strict privacy requirements for for such a secure cloud information usage framework. Among different multi-keyword semantics, they choose the proficient similarity measure of “coordinate matching,” i.e., whatever number matches as possible, to capture the relevance of data documents to the search query. They further utilize “inner product similarity” to quantitatively assess such similarity measure. They first propose a fundamental thought for the MRSE based on secure inner product computation, and after that give two essentially enhanced MRSE plans to accomplish different stringent privacy requirements in two distinctive threat models.

Sun et al. [6], authors present a privacy-preserving multi-keyword text search (MTS) theme with similarity-based ranking to address this issue. To support multi-keyword search and search result ranking, they propose to create the search index based on term frequency and also the vector area model with cosine similarity live to attain higher search result accuracy. To enhance the search efficiency, they propose a tree-based index structure and numerous adaption ways for multi-dimensional (MD) algorithmic rule so the sensible search efficiency is way higher than that of linear search. To further enhance the search privacy, they propose two secure index schemes to fulfill the stringent privacy needs under strong threat models, i.e., known cipher text model and known background model.

Chen et al. [7], a hierarchical clustering technique for cipher text search within a big data environment is proposed. The proposed approach clusters the documents based on the

minimum similarity threshold, and after that segments the resultant clusters into sub-clusters until the constraint on the maximum size of cluster is reached. In the search phase, this methodology can achieve a linear computational complexity against exponential size of document collection. In addition, retrieved documents have a better relationship with each other than traditional methods.

3. PROPOSED APPROACH

3.1 Problem statement

To design a system that provides relevant rank results by using hierarchical clustering index and data duplication check is performed to reduce memory overhead and searching time. It provides more security by user-revocation approach.

3.2 Proposed System overview

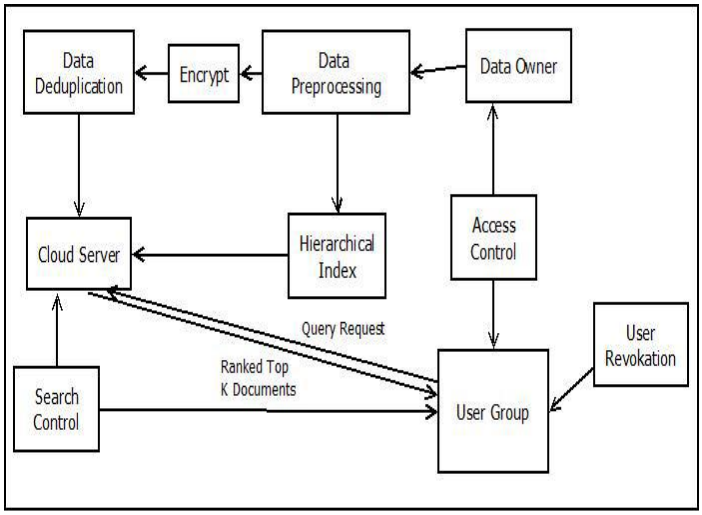

Figure 1. Proposed System Architecture

The propose system consist of three entities data owner, the data user and the cloud server. The data owner is responsible for collecting documents, building document index and outsourcing them in an encrypted format to the cloud server. Aside from that, the data user needs to get the authorization from the data owner before getting to the information. The cloud server gives a huge storage space and the computation resources required by cipher text search. Upon receiving a legal request from the data user, thecloud server searches the encrypted index, and sends back top-k documents that are most likely to match users query. The number k is appropriately chosen by the data user. This framework goes for protecting data from leaking information to the cloud server while improving the efficiency of cipher text search. In this model, both the data owner and the data user are trusted, while the cloud server is semi-trusted.

The proposed system consists of different modules the main modules are listed below:

- Cloud Server:
  Cloud server is the storage medium on which the data is stored. Data stored on the cloud is in encrypted form by which only authenticated users are able to access the data in this way security is provided for data stored on cloud

- Data Owner:
  Data owner are the user which uses the cloud storage to store the data on cloud for the sharing purpose with the different user present in the group. He is the one responsible for encryption of data while storing it on cloud.

- User Group:
  This is the group of users who are willing to use the data stored on the cloud. They have made a request in order to access the data.

- Access control module:
  At the time when user enters in the group or leaves the group this module is responsible for providing the access keys when one user enters in the group also responsible for the key revocation at the time when the user leaves the group.

- Search Control:
  This module is brought in to action when user searches for specific keywords on the cloud. This module is able to retrieve the files with the top k ranked documents acceding to the user search.

- User Revocation and Data Deduplication:

To provide the security this system used user revocation approach. To enhance the searching speed data deduplication check is performed at client side. In this approach the repeated or duplicate data is removed and minimize memory overhead and searching time.

Steps of the proposed system are as follows:

Document Vector Generation

Initially User browse the dataset file, perform the operation of stemming, stop word which remove the redundant data, and improve the performance of the system,

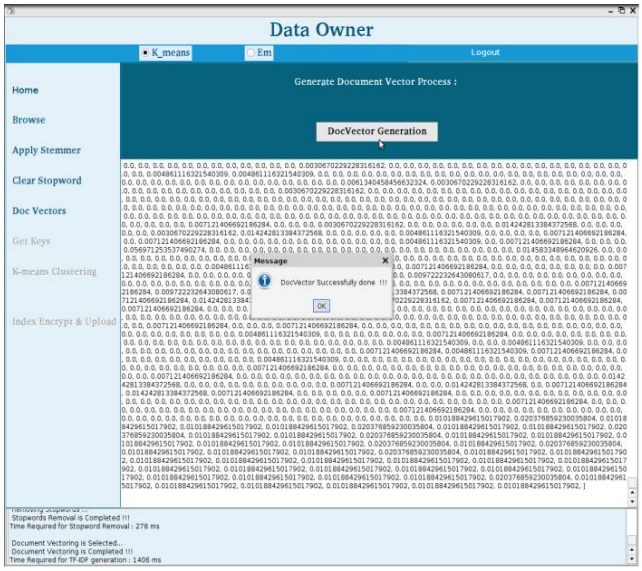

Figure -2: Document Vector Generation

K-Means Clustering

After document vector generation, clustering process is performed. User enter the number of clusters and perform the k-means clustering operation

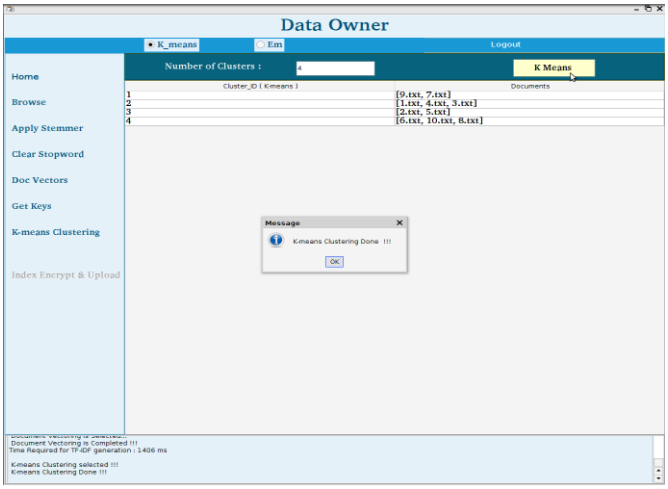

Figure-3: K-means Clustering

- Check Deduplication

System generates the indexing file, encrypt index file, encrypt files and generate the hash value by using SHA1 algorithm. And final system checks for the deduplication file.

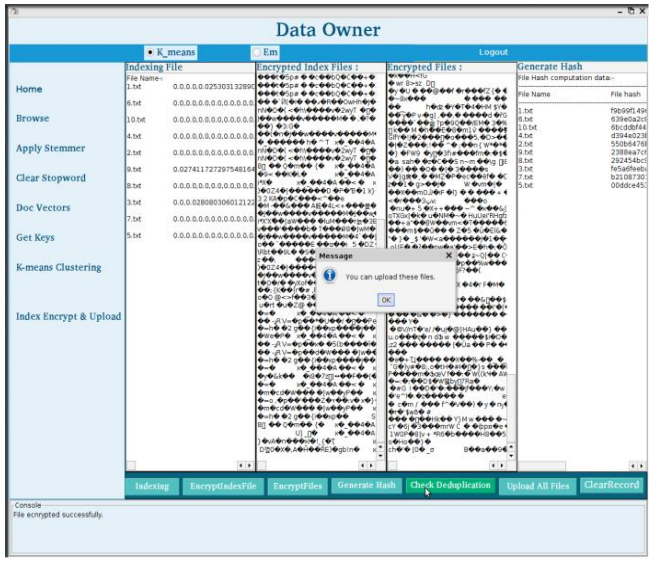

Figure -4: Duplication Check

EM Clustering

As EM clustering is our contribution work, we test the performance of the system, by implementing the system by using EM clustering algorithm

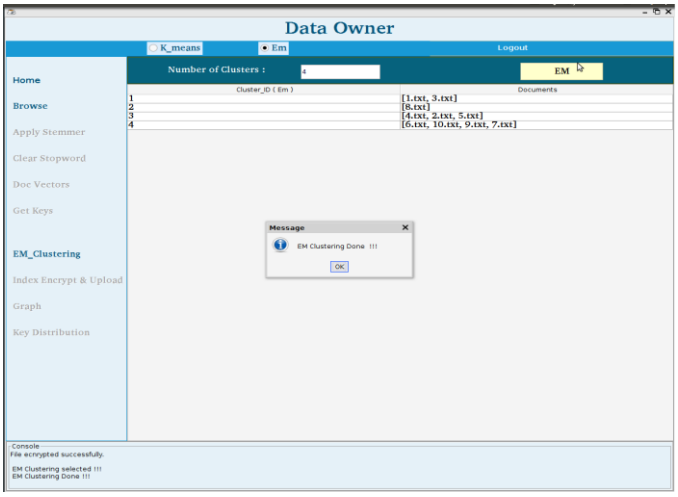

Figure-5 EM Clustering

3.3 ALGORITHM

In MRSE-HCI architecture 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. This architecture mainly consists of following algorithms.

- $\text{Keygen}\left(1^{l(n)}\right) \rightarrow (\text{sk}, \text{k})$: It is used to generate the secret key to encrypt index and documents.
- $\text{Index}(\text{D}, \text{sk}) \rightarrow \text{I}$: 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.

- $Enc(D,k) \rightarrow E$: The document collection is encrypted by a symmetric encryption algorithm which achieves semantic security.
- $Trapdoor(w, sk) \rightarrow T_w$: It generates encrypted query vector T_w with users input keywords and secret key.
- $Search(T_w, I, k_{top}) \rightarrow (I_w, E_w)$: In this phase, cloud server compares trapdoor with index to get the top-k retrieval results.
- $Dec(E_w, k) \rightarrow F_w$ he returned encrypted documents are decrypted by the key generated in the first step.

Algorithm 2: Deduplication Checking

1. input the initial set of k cluster Centers C
2. set the threshold TH min
3. while k is not stable
4. generate a new set of cluster centers CΘ by k-means
5. for every cluster centers CΘλ
6. get the minimum relevance score: min (Si)
7. if the min (Si) <THmin
8. add a new cluster center: k = k + 1
9. go to while
10. until k is steady

Algorithm 3: K-means Algorithm

1. db Hash=getting files hash from the client side database which are successfully uploaded to the server
2. File Hash=users chunked file's hash
   H(New file) = h
   H(Old n chunks) = hn
   Compare h and hn
   If H(New chunk) == H(Old n files)
   Chunk is duplicate and refuses it to store on public server
   Else
   Chunk is not duplicate and allowed to store on public server

Algorithm 4: SHA-1 Algorithm

Steps of SHA-1

Step 1: Append Padding Bits Message is"padded" with a 1 and as many 0's as necessaryto bring the message length to 64 bitsfewer than an even multiple of 512.

Step 2: Append Length 64 bits are appendedto the end of the padded message. These bitshold the binary format of 64 bits indicating thelength of the original message.

Algorithm 5: User Revocation

Input: User Name

Output: User terminated from system

Retrieve user List

Select user from the retrieve user list

Closed all operation from system (Selected User).

Then put that selected user in revoked list.

4. RESULTS AND DISCUSSION

In this section discussed the experimental result of the proposed system.

4.1 Time Comparison Graph for Clustering

Following chart 1 shows the time comparison graph of the different clustering algorithm. Existing system used k means algorithm for the clustering process, proposed system used the EM clustering algorithm for the clustering process. As from the following graph it is conclude that time required for completing the process of clustering by using the K-means algorithm is more than the EM clustering algorithm. Hence by using the EM clustering algorithm performance of the system is improved.

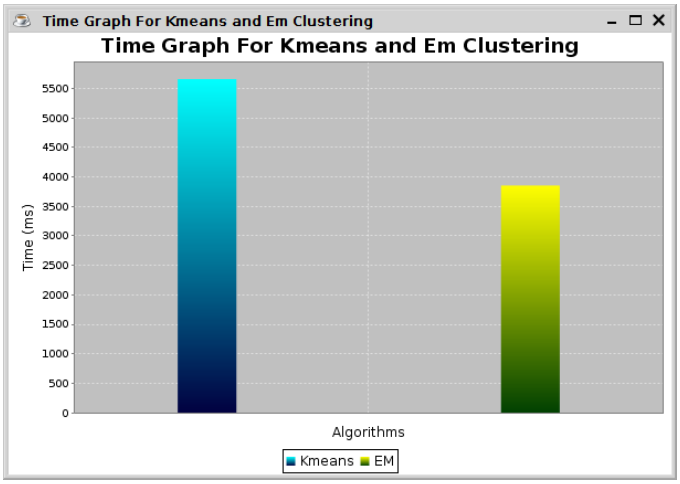

Chart -1: Time Comparison graph for clustering

4.2 Time Comparison Graph for Deduplication

As if the duplicate file exist in the system, the performance of the system may be low, therefore in the proposed system we implement the concept of deduplication. In the following chart 2 shows the comparison of existing system without using the concept of deduplication and proposed system with using the concept of deduplication. From the figure it is conclude that time required for the system with deduplication is less than the time required for the system without deduplication.

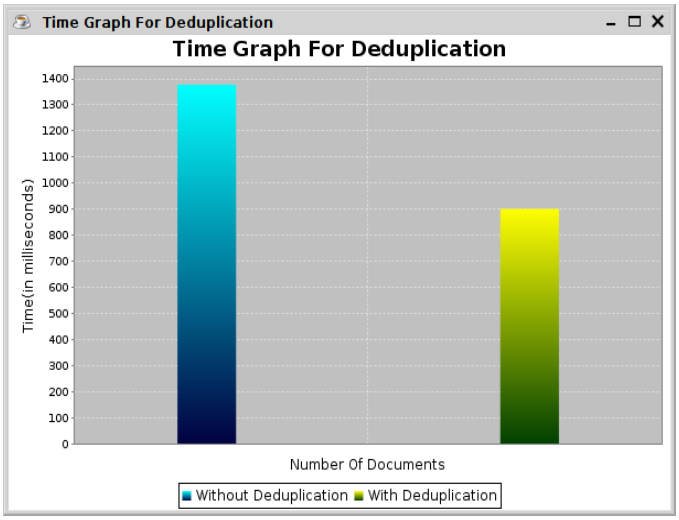

Chart 2. Time Comparison graph for deduplication

4.3 Time Comparison Graph for Searching keyword

The search accuracy can measure the clients fulfillment. The Retrieval accuracy is identified with two components: the relevance amongst reports and the queries and the relevance of documents between each other.

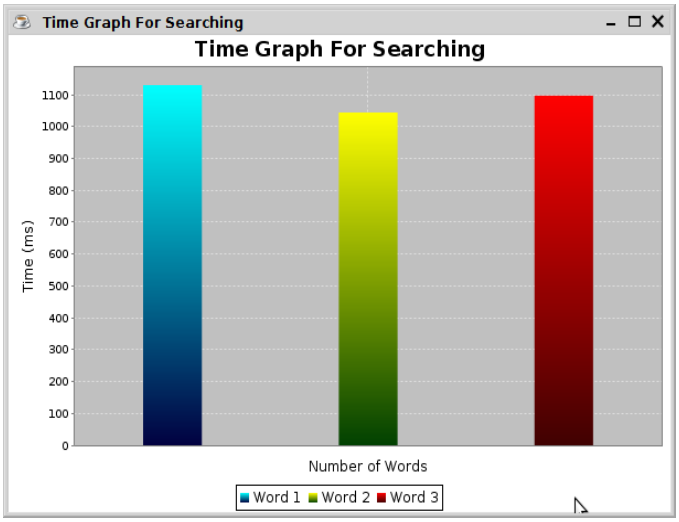

Chart-3. Time Comparison graph for searching Keyword

4.4 Relevance of Document

From the chart4, we can observe that the relevance of retrieved documents in the MRSE-with clustering is almost twice as many as that in the MRSE-HCI, which means retrieved documents generated by MRSE-with clustering are much closer to each other.

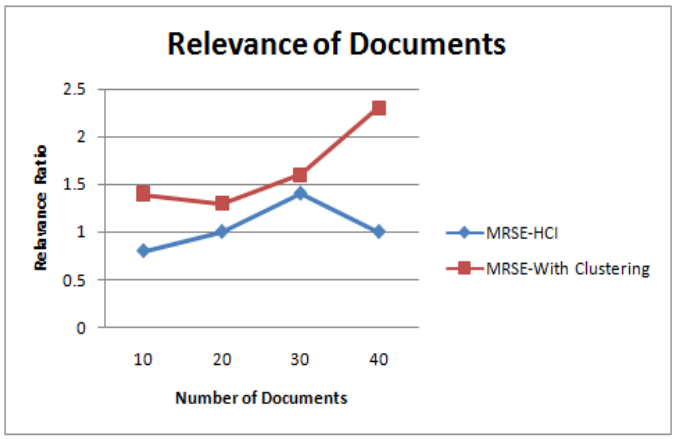

Chart-4 : Relevance of Document

4.6 Search Time Graph

Chart5 describes search efficiency using the different size of document set with unchanged dictionary size, number of retrieved documents and number of query keywords,

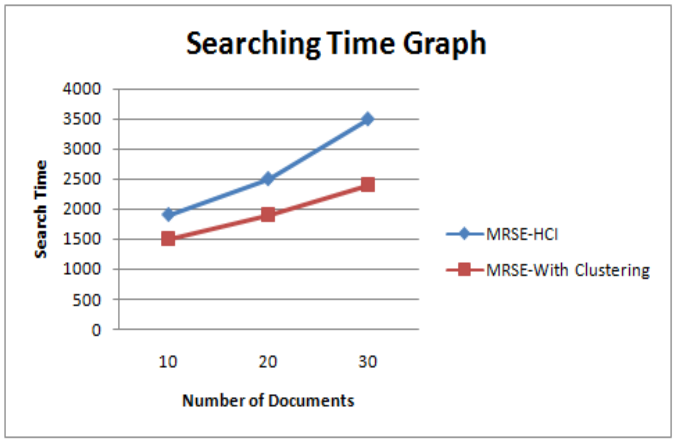

Chart. 5: Search Time Graph

5. CONCLUSIONS

This system examined cipher text search in the circumstances of cloud storage. Also discuss the issues 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. The existing MRSE-HCI architecture adapts the requirements of data explosion, online information retrieval and semantic search. The experiment result proves that the proposed architecture not only properly solves the multi-keyword ranked search problem, but also brings and improvement in search efficiency, rank security, and the relevance between retrieved documents The proposed system enhance the system performance by implementing user revocation method where user group revocate Also system reduces the memory overhead and enhances

searching speed by implementing data deduplication approach where duplicate data is removed.

REFERENCES

[1]. Chen, Chi, et al."An efficient privacy-preserving ranked keyword search method."IEEE Transactions on Parallel and Distributed Systems 27.4 (2016): 951-963.

[2] S. Grzonkowski, P. M. Corcoran, and T. Coughlin, "Security analysis of authenticationprotocols for next-generation mobile and CE cloud services," in Proc. IEEEInt. Conf. Consumer Electron., 2011, Berlin, Germany, 2011, pp. 83-87.

[3] C. Wang, N. Cao, K. Ren, and W. J. Lou, ``Enabling secure and efficient rankedkeyword search over outsourced cloud data," IEEE Trans. Parallel Distrib. Syst., vol. 23, no. 8, pp. 1467-1479, Aug. 2012.

[4] H. Pang, J. Shen, and R. Krishnan, "Privacy-preserving similarity based text retrieval,"ACM Trans. Internet Technol., vol. 10, no. 1, pp. 39, Feb. 2010.

[5] N. Cao, C.Wang, M. Li, K. Ren, andW. J. Lou, "Privacy-preserving multi-keywordranked search over encrypted cloud data," in Proc. IEEE INFOCOM,Shanghai,China, 2011, pp. 829-837.

[6] W. Sun, B.Wang, N. Cao, M. Li, W. Lou, Y. T. Hou, and H. Li, "Privacy-preservingmulti-keyword text search in the cloud supporting similarity-based ranking,"inProc. 8th ACM SIGSAC Symp.Inform.,Comput. Commun.Security, Hangzhou,China, 2013, pp. 71-82.

[7] C. Chen, X. J. Zhu, P. S. Shen, and J. K. Hu, "A hierarchical clustering methodFor big data oriented cipher text search," in Proc. IEEE INFOCOM, Workshop onSecurity and Privacy in Big Data, Toronto, Canada, 2014, pp. 559ˆa564.

[8] Cash, David, et al. "Dynamic Searchable Encryption in Very-Large

[9] Yanzhu Liu, Zhi Li, Wang Guo and Wu Chaoxia, "Privacy-preserving multikeyword ranked search over encrypted big data," Third InternationalConference on Cyberspace Technology (CCT 2015), Beijing, 2015, pp.1-3.

[10] Ching-Yang Tseng, ChangChun Lu and Cheng-Fu Chou, "Efficient privacypreserving multi-keyword ranked search utilizing document replication and partition," 2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, 2015, pp.

[11] C. Wang, N. Cao, K. Ren, and W. J. Lou, "Secure ranked keyword search over encrypted cloud data"The 30th International Conference on Distributed Computing Systems (ICDCS'10), Genoa, Italy, June 21-25, 2010.

[12] S. C. Yu, C. Wang, K. Ren, and W. J. Lou, "Achieving Secure, Scalable, and Fine-grained Data Access Control in Cloud Computing," in Proc. IEEE INFOCOM, San Diego, CA, 2010, pp. 1-9.