

# Hierarchical Clustering Over Rank Keyword Search with Secure and Competent Outlook

<sup>1</sup>Ms. Shumaila Tajreen, <sup>2</sup>Dr. G.S.S Rao

<sup>1</sup>M.Tech(CSE), <sup>2</sup>Professor & HOD

<sup>1,2</sup>Nawab ShahAlam khan College of Engineering and Technology, Hyd

Email: [shumailatajreen@gmail.com](mailto:shumailatajreen@gmail.com) , [profgssrao@gmail.com](mailto:profgssrao@gmail.com)

**Abstract:** As the volume of data in the data center is experiencing a tremendous growth, so vendor of data prefers to outsource sensitive and important documents for the purpose of privacy conserving. The documents are stored in encrypted format so it is essential to develop efficient search cipher text technique. In the process of encryption relationship between document is concealed which leads to perform deterioration. In this paper, a quality hierarchical clustering (QHC) method is proposed to support searching mechanism and to meet fast searching within cloud environment. In this paper multi-keyword ranked search, hierarchical clustering index (MRSE-HCI) architecture is used. For search result verification, least hash subtree is used. The proposed method has several advantages over traditional method in document's retrieval and rank privacy.

**Keywords:** Cloud Computing, Hierarchical Clustering, Security, Ciphertext search, multi keyword search, ranked search.

## 1.Introduction

As we step into a digital era, terabyte's of data is generated worldwide per day, so any data owner who wants to outsource their data need to be step up in a cloud computing. Organizations and enterprises with a large amount of data prefer to outsource data in order to reduce data management cost and storage facility. Although cloud service provider's (CSP) claim for their services regarding security, privacy measure security and privacy are major obstacles preventing for wider acceptance. A traditional way of data encryption makes server-side utilization such as searching on encrypted data becomes a challenging task. Therefore, proposing a method which can maintain and utilize this relationship to speed the search phase is desirable. Cloud computing is the open source platform, broad network access, on-demand capability, rapid elasticity- are the several advantages of cloud computing but security and privacy are the major issues. In existing solutions, data are stored in plain text due to which data is vulnerable to attacks.



Fig. 1. Architecture of cipher text search

In this paper, 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.

## 2.Literature Survey

With the advantage of storage as a service, many enterprises are moving their valuable data to the cloud since it costs less, easily scalable and can be accessed from anywhere anytime [1]. The trust between cloud

user and provider is paramount. Here security as a parameter is used to establish trust. Cryptography is one way of establishing trust. Searchable encryption is a cryptographic method to provide security. In literature, many researchers have been working on developing efficient searchable encryption schemes. This paper explores some effective cryptographic techniques based on data structures like CRSA and B-Tree to increase the level of security. It tried to implement the search on encrypted data using Azure cloud platform [1]. Cloud computing is generating a lot of interest to provide absolute for data outsourcing and high-quality data services. More and more institutions, organizations and corporations are exploring the possibility of having their applications, data and their IT assets in cloud [2]. As the data and their cloud's size increases searching of the relevant data is expected to be a challenge. To overcome this challenge, search index is created to aid in more rapidly search. However, search Index creation and computation have been complex and time-consuming, leading to cloud downtime thereby hindering the swiftness in reacting to data request for mission-critical requirements. The focus of this paper is to explain how reusability of search index is helping to reduce the complexity of search index computation. The search index is proposed to be created using parameters like similarity relevance, user ranking, and scheme robustness. User ranking helps to guarantee a keyword is used frequently in the uploaded data [2]. The proposed system defined that the reusability of search index concept reduces cloud consuming time while maintaining the security using searchable symmetric encryption (SSE). The file requested by client is fetched from the cloud, using Two- round searchable encryption (TRSE) scheme that supports multi-keyword retrieval [2].

Nowadays, more and more people are motivated to outsource their local data to public cloud servers for great convenience and reduced costs in data management. But in consideration of privacy issues, sensitive data should be encrypted before outsourcing, which obsoletes traditional data utilization like keyword-based document retrieval. This paper presents a secure and efficient multi-keyword ranked search scheme over encrypted data, which additionally supports dynamic update operations like deletion and insertion of documents [3]. Specifically, we construct an index tree based on vector space model to provide multi keyword search, which meanwhile supports flexible update operations. Besides, cosine similarity measure is utilized to support accurate ranking for seek result. To improve search efficiency, we further propose a search algorithm based on “Greedy Depth first Traverse Strategy”. Moreover, protecting the search privacy, propose a secure scheme to meet various privacy requirements in the known cipher text threat model. Experiments on the real

world dataset show the effectiveness and efficiency of projected scheme [3].

### 3. Our Contribution

The problem of maintaining the close relationship between different plain documents over an encrypted domain has been investigating and proposes a clustering method to solve this problem. We design a search strategy to enhance the rank privacy. This search strategy adopts the backtracking algorithm upon the above clustering method, by using backtracking algorithm we solve any query where it occurs. 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.

### 4. Definitions and Background

#### 4.1 Threat Model

The adverse ability is concluded in two threat models. Known ciphertext model: In this model, Cloud server can get encrypted document collection, encrypted query keywords, and encrypted data index. Known background model: In this model, cloud server knows more information than that in known ciphertext model. Statistical background information of dataset, such as the document frequency and term frequency information of a specific keyword, can be used by the cloud server to launch a statistical attack to infer or identify a specific keyword in the query which further reveals the plain-text content of documents.

#### 4.2 Design Goals

Search efficiency. The time complexity of search time of the MRSE-HCI is less where scheme needs to be logarithmic against the size of data collection in order to deal with the explosive growth of document size in big data scenarios.

Retrieval accuracy. Retrieval precision is related to two factors: the relevance between the query and the documents in outcome set, and the relevance of documents in the result set.

The integrity of the search results the correctness, completeness, and freshness of the document should be maintained.

### 5. System Architecture And Algorithm

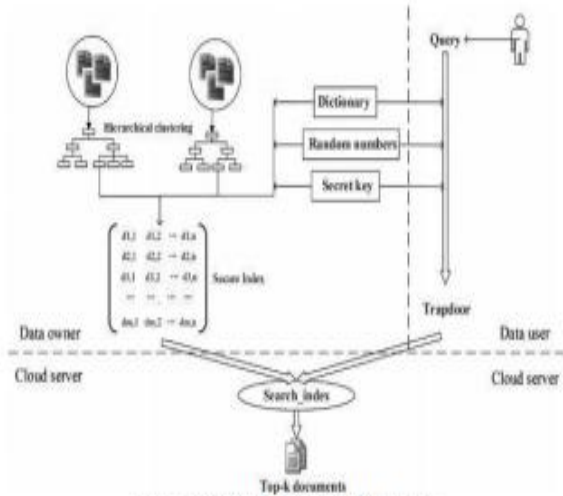


Fig. 2. MRSE-HCI architecture

The MRSE-HCI architecture is depicted by Fig.2. where the data owner is responsible for collecting documents, building document index and outsourcing them in an encrypted format to the cloud server. The cloud 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 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. The data user needs to get the authorization from the data owner before access to the data.

Until now many hierarchical clustering methods have been proposed but all this method is not comparable to the partition clustering method, K-mean and KMedoid are popular clustering algorithms but the size of k is fixed here. So we proposed a quality hierarchical clustering algorithm based on dynamic k-means.

Algorithmic Dynamic k-means 1, input the initial set of k cluster C 2, set the threshold THmin 3, while k s not stable 4, generate a new set of cluster center C0 by k-means 5, for every cluster center C0, I 6, get the minimum relevance score: min(Si) 7, if the min(Si).

Every cluster is checked whether its size exceeds TH or not. If the size exceeds the cluster will split into child cluster which is formed by dynamic k-means this procedure will be repeated until the entire cluster meet the requirement of maximum cluster size.

Algorithm Quality Hierarchical Clustering (QHC)

1, input document and set the size threshold TH

- 2, build cluster ser C0 in 1<sup>st</sup> level by dynamic k-means
- 3, while their new cluster set Ci
- 4, for every cluster Ci,j
- 5, if the size Ci,j is bigger than TH
- 6, split this cluster into sub-cluster Ci+1
- 7, until all clusters, match the size constraint

The retrieved document has a prospect to be wrong because of unstable network and the data may damage due to hardware or software failure, so verifying the authenticity of seeks result is acritical issue in cloud environment. Therefore, the minimum hash subtree is designed to verify the correctness and freshness of search result.

Algorithm building minimum hash sub-tree (MHST)

- 1, build hast 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 (idj)
- 8, goto the upper level
- 9, calculate tree root's hash value:
- 10, calculate the signature of hash value

## 6. Conclusion

The problem of maintaining the semantic relationship between different plain documents has been explored and given the design method to enhance the performance of the semantic search. For the adaptation to the requirements of data explosion, online information retrieval, and semantic search, MRSE-HCI architecture has been proposed. Accordingly, a verifiable mechanism is proposed for the guarantee of correctness and completeness of search results. In addition, the search efficiency and security under two popular threat models have been analyzed. An experimental platform is built to evaluate the search efficiency, accuracy, and rank security. 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

## References

- [1]. Dynamic Multi-Keyword Ranked Searchable Security Algorithm Using CRSA and B-Tree Prasanna B T#1, C B Akki\*2 #Department of ISE, EPCET Associate Professor, Bengaluru, INDIA-560049 \*Department of ISE, SJBIT Professor, Bengaluru, INDIA-560060
- [2]. Reusability of Search Index over Encrypted Cloud Data on Dynamic update Kavitha R1, R J Poovaraghan2 Student, M.Tech, SRM University, Chennai, India1

Assistant Professor (OG), Department of Computer Science, SRM University, Chennai, India2.

[3]. Dynamic Multi-keyword Top-k Ranked Search over Encrypted Cloud Data Xingming Sun, Xinhui Wang, Zhihua Xia, Zhangjie Fu and Tao Li Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing, 210044.

[4]. R. Ostrovsky, D. Boneh, G. Di Crescenzo, and G. Persiano, “Public key encryption with keyword search,” in *Advances in Cryptology-Eurocrypt 2004*, 2004, pp. 506–522.

[5]. R. Curtmola, J. Garay, S. Kamara and R. Ostrovsky “Searchable symmetric encryption: Improved definitions and efficient constructions”*Proc. 13th ACM Conf. Comput. Commun. Secur.*, pp. 79-88, 2006.

[6]. G. Craig “Fully homomorphic encryption using ideal lattices”*Proc. 41st Annu. ACM Symp. Theory Comput.*, vol. 9, pp. 169-178, 2009.

[7]. 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.

[8]. Q.Wang et al, K. Ren, C.Wang, “Security challenges for the public cloud,” *IEEE Internet Computing*, 2012, vol. 16, no. 1, pp. 69–73.

[9]. S. Kamara, C. Papamanthou and T. Roeder “Dynamic searchable symmetric encryption”*Proc. Conf. Comput. Commun. Secur.*, pp. 965-976, 2012.

[10]. 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.

[11]. D. Cash, S. Jarecki, C. Jutla, H. Krawczyk, M. Rosu and M. Steiner “Highly-scalable searchable symmetric encryption with support for Boolean queries”*Proc. Adv. Cryptol.*, pp. 353-373, 2013.

[12]. K. Ren, C. Wang, N. Cao, and W. Lou, “Enabling secure and efficient ranked keyword search over outsourced cloud data,” *Parallel and Distributed Systems*, *IEEE Transactions on*, 2012, vol. 23, no. 8, pp. 1467–1479.

[13]. S. Jarecki, D. Cash, J.Jaeger, C. Jutla, M. C. Rosu, and M. Steiner, “Dynamic searchable encryption in very large databases: Data structures and implementation”, vol. 14, 2014,

[14]. 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.

[15]. 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.

[16] A.Raghavendra Praveen Kumar, K.Tarakesh, and U.Veeresh ,” A Secure and Dynamic Multi Keyword Ranked Search Scheme over encrypted.” *International*

*Journal of Computer Engineering in Research Trends.*, vol.2, no.12, pp. 1137-1141, 2015.

[17] Mr. Rahul Hon, and Mrs. N.Sujatha,” A Novel Framework to Measure the Degree of Difficulty on Keyword Query Routing.” *International Journal of Computer Engineering in Research Trends.*, vol.3, no.6, pp. 314-320, 2016.

[18] Kallem Rajender Reddy, and Y.Sunitha,” A Novel Framework to Measure the Degree of Difficulty on Keyword Query Routing.” *International Journal of Computer Engineering in Research Trends.*, vol.2, no.9, pp. 640-645, 2015.

[19] Vadla Jhansi Rani, and K.Samson Paul,” Secure Multi Keyword Dynamic Search Scheme Supporting Dynamic Update.” *International Journal of Computer Engineering in Research Trends.*, vol.4, no.8, pp. 356-360, 2017.

[20] Mr. M. Veerabrahma Chary and Mrs.N.Sujatha,” A Novel Additive Multi-Keyword Search for Multiple Data Owners in Cloud Computing.” *International Journal of Computer Engineering in Research Trends.*, vol.3, no.6, pp. 308-313, 2016.