e-ISJN: A4372-3114
p-ISJN: A4372-3115

ISSN: 2321-7782 (Online)
ISSN: 2347-1778 (Print)
Impact Factor: 7.327

Volume 8, Issue 5, May 2020

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

# *Privacy-Preserving Multi-keyword Top-k Ranked Similarity Search Over Encrypted Data*

**Namrata G. Choudhary**[1]
ME II nd Year Student & Researcher
Department of Computer Science & Engineering
Matoshri Pratishthan Group of Institutions,
Nanded, Maharashtra, India

**Ashok Namdeo Kamthane**[2]
Associate Professor
Department of Computer Science & Engineering,
Matoshri Pratishthan Group of Institutions,
Nanded, Maharashtra, India

***Abstract:*** ***Cloud computing provides the facility to store and manage data remotely. The volume of information is increasing per day. The owners choose to store the sensitive data on the cloud storage. To protect the data from unauthorized accesses, the data must be uploaded in encrypted form. Due to large amount of information is stored on the cloud storage; the association between the documents is hiding when the documents are encrypted. It is necessary to design a search technique which gives the results on the basis of the similarity values of the encrypted documents. In this paper a cosine similarity clustering method is proposed to make the clusters of similar documents based on the cosine values of the document vectors. We also proposed a MRSE-CSI model used to search the documents which are in encrypted form. The proposed search technique only finds the cluster of documents with the highest similarity value instead of searching on the whole dataset. Processing the dataset on two algorithms shows that the time needed to form the clusters in the proposed method is less. When the documents in the dataset increases, the time needed to form clusters also increases. The result of the search shows that increasing the documents also increases the search time of the proposed method.***

***Keywords:*** ***Cloud computing, multi-Keyword search, cosine similarity clustering, encrypted data.***

## I. INTRODUCTION

Cloud computing becomes popular as it provides huge storage space and high quality services. The large amount of data is created per day. It is a difficult task for the owner of the data to store and manage this large amount of data. To overcome this difficulty, the data owners can store their data on the cloud server to use the on demand applications and services from shared resources [1]. The cloud server providers agreed that their cloud service is armed with strong security constraints though security and privacy are major hindrances which avoid the use of cloud computing services [2]. To protect the sensitive data on the cloud server from unauthorized users, the data owners may encrypt the documents and uploads to cloud server [3]. In the earlier various strong cryptography methods were used to design the search techniques on the cipher text [4], [5], [6]. These techniques needs many operations and require large amount of time. So these techniques are not suitable for big data where information volume is huge. The property of a document depends on its association The results of search returned to the users may contain damaged information due. Thus a mechanism should be given to users to check the accuracy of the search results. The proposed architecture of search technique is based on the cosine similarity clustering which maintain the association between plain text and encrypted text to improve the efficiency of search.

## II. LITERATURE SURVEY

Chi Chen and Xiaojie Zhu [7] used a hierarchical clustering method to maintain the close relationship between plain documents and encrypted documents to increase search efficiency within a big data environment. They also used a coordinate

matching technique [8] to measure the relevance score between query and document. They did a model for the efficient multi-keyword ranked search and maintain the privacy of documents, rank security and relevance between retrieved documents. Jiadi Yu and Peng Lu [9] focused on the problems of the cipher text search using Searchable Symmetric Encryption (SSE) [10], [11]. This SSE technique helps data users to retrieve the documents over the encrypted documents. In Two Round Searchable Encryption (TRSE), they used the similarity relevance concept to solve the privacy issues in searchable encryption. They also showed server side ranking according to order preserving encryption (OPE). N. Cao, C. Wang and M. Li [12] used "inner product similarity" concept which can find the similarity measure of the information and the keywords of search. Ruksana Akter, Yoojin Chung [13] defined an evolutionary approach based on cosine similarity clustering. A document vector is used to create the index of every document. The cosine values between the document vectors are calculated. Clusters of the most relevant documents are formed on the basis of the cosine values. Another good feature of their work.

**DISADVANTAGES OF EXISTING SYSTEM:**

1. Applying these approaches for data encryption usually cause tremendous cost in terms of data utility, which makes traditional data processing methods that are designed for plaintext data no longer work well over encrypted data.
2. Most of these methods cannot meet the high search efficiency and the strong data security simultaneously, especially when applying them to big data encryption poses great scalability and efficiency challenges.
3. But the cost of search remains high and the time complexity of creating trapdoor is high.

## III. PROPOSED SYSTEM

We define and solve the challenging problem of privacy-preserving multi-keyword ranked search over encrypted cloud data (MRSE), and establish a set of strict privacy requirements for such a secure cloud data utilization system to become a reality. Among various multi-keyword semantics, we choose the efficient principle of "coordinate matching", i.e., as many matches as possible, to capture the relevance of data documents to the search query. Specifically, we use "inner product similarity", i.e., the number of query keywords appearing in a document, to quantitatively evaluate such similarity measure of that document to the search query. In cloud computing, data owners may share their outsourced data with a number of users, who might want to only retrieve the data files they are interested in. One of the most popular ways to do so is through keyword-based retrieval. Keyword- based retrieval is a typical data service and widely applied in plaintext scenarios, in which users retrieve relevant files in a file set based on keywords. However, it turns out to be a difficult task in cipher text scenario due to limited operations on encrypted data. Besides, to improve feasibility and save on the expense in the cloud paradigm, it is preferred to get the retrieval result with the most relevant files that match users' interest instead of all the files, which indicates that the files should be ranked in the order of relevance by users' interest and only the files with the highest relevance's are sent back to users. on-demand high-quality applications and services from a shared pool of Fig.1.Proposed Architecture configurable computing resources [2], [3]. Its great flexibility and economic savings are motivating both individuals and enterprises to outsource their local complex.

 

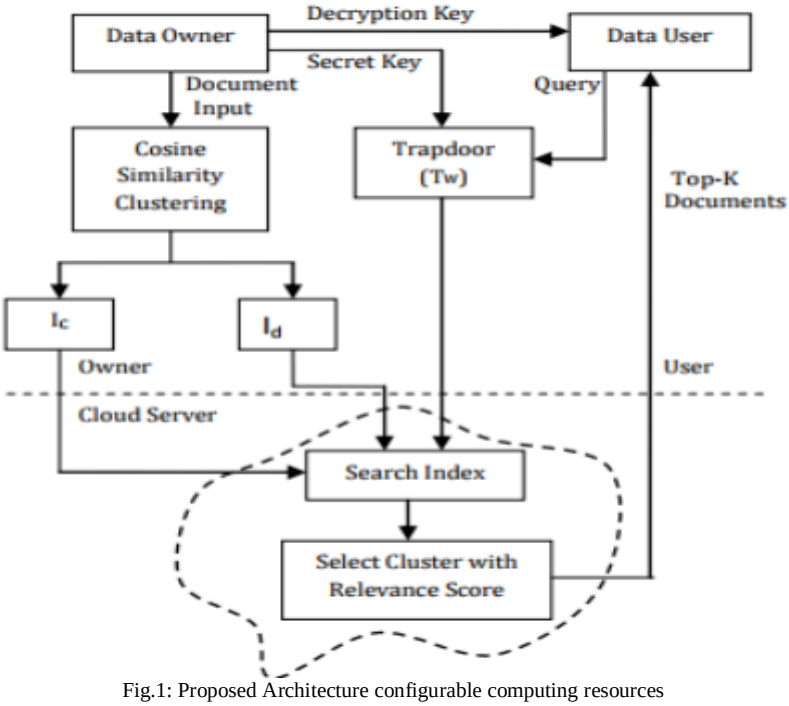


Fig.1: Proposed Architecture configurable computing resources

## IV. ANALYSIS


Fig.2.Output View A simulation toolkit

Fig.2.Output View a simulation toolkit enables modeling and simulation of Cloud computing systems and application provisioning environments. The Cloud Sim toolkit supports both system and behavior modeling of Cloud system components such as data centers, virtual machines (VMs) and resource provisioning policies. It implements generic application provisioning techniques that can be extended with ease and limited effort. Currently, it supports modeling and simulation of Cloud computing environments consisting of both single and inter-networked clouds (federation of clouds). Moreover, it exposes custom interfaces for implementing policies and provisioning techniques for allocation of VMs under inter-networked Cloud computing scenarios. In this module we are creating cloud users and datacenters and cloud virtual machines as per our requirement. The term instance type will be used to differentiate between VMs with different hardware characteristics. The Retrieval phase involves Trapdoor Gen, Score Calculate, and Rank, in which the data user and the cloud server are involved. As

 

a result of the limited computing power on the user side, the computing work should be left to server side as much as possible. Meanwhile, the confidentiality privacy of sensitive information cannot be violated. The ranking should be left to the user side while the cloud server still does most of the work without learning any sensitive information. In the first experiment, document search time is calculated. Five different number of dataset sizes are chosen in the experiment to show the effect on the efficiency of the search results. From chart 1, we can see that the time needed to search the documents increases when the size of dataset increases. Compared with the previous related work [7] time needed to search the documents is less.

## V. CONCLUSION

We motivate and solve the problem of secure multi keyword top-k retrieval over encrypted cloud data. In this work, a new framework is proposed for the problem of multi-keyword ranked search over encrypted cloud data, and to establish a variety of privacy requirements. Among various multi- keyword semantics, the efficient similarity measure is "coordinate matching", i.e., as many matches are possible, to effectively capture the relevance of outsourced documents to the query keywords, and use "inner product similarity" to quantitatively evaluate such similarity measure. For meeting the challenge of supporting multi-keyword semantic without privacy breaches, MRSE framework is proposed using secure inner product computation. Thorough analysis investigating privacy and efficiency guarantees of proposed schemes is given, and experiments on the real-world dataset shows our proposed scheme introduces low overhead on both computation and communication.

## References

1. R. Curtmola, J.A. Garay, S. Kamara, and R. Ostrovsky, "Searchable Symmetric Encryption: Improved Definitions and Efficiet Constructions," Proc. ACM 13th Conf. Computer and Comm. Security (CCS), 2006.
2. L.M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A Break in the Clouds: Towards a Cloud Definition," ACM SIGCOMM Comput. Commun. Rev., vol. 39, no. 1, pp. 50-55, 2009.
3. N. Cao, S. Yu, Z. Yang, W. Lou, and Y. Hou, "LT Codes- Based Secure and Reliable Cloud Storage Service," Proc. IEEE INFOCOM, pp. 693-701, 2012.
4. D. Song, D. Wagner, and A. Perrig, "Practical Techniques for Searches on Encrypted Data," Proc. IEEE Symp. Security and Privacy, 2000.
5. E.-J. Goh, "Secure Indexes," Cryptology ePrint Archive, http://eprint.iacr.org/2003/216. 2003.
6. Y.-C. Chang and M. Mitzenmacher, "Privacy Preserving Keyword Searches on Remote Encrypted Data," Proc. Third Int'l Conf. Applied Cryptography and Network Security, 2005.
7. R. Curtmola, J.A. Garay, S. Kamara, and R. Ostrovsky, "Searchable Symmetric Encryption: Improved Definitions and Efficient Constructions," Proc. 13th ACM Conf. Computer and Comm. Security(CCS '06), 2006.

 

## AUTHOR(S) PROFILE

**Ms. Namrata G. Choudhary,** has received BE degree in Computer Science Engineering from MGM's college of Engineering Nanded, Maharashtra, India.in 2013. She is currently pursuing Master of Engineering in computer science and Engineering with dissertation topic on privacy preserving Multi-keyword Top-k search over encrypted data.

**Ashok Namdeo Kamthane,** Qualification: ME Electronics Experience- 37 years.