ZKG INTERNATIONAL ISSN: 2366-1313

A Fast and Accurate Privacy-Preserving Multi-keyword Top-k Retrieval Scheme over Encrypted Cloud Data

[1]Mrs. B RAJANI, [2]K POOJITH GOUD, [3]J SINDHUJA, [4]K PRANAY, [5]K SURESH

[1](Assistant Professor) ,**CSE.** Teegala Krishna Reddy Engineering College Hyderabad

[2345]B,tech scholar ,**CSE.** Teegala Krishna Reddy Engineering College Hyderabad

ABSTRACT

Cloud computing provides individuals and enterprises massive computing power and scalable storage capacities to support a variety of big data applications in domains like health care and scientific research, therefore more and more data owners are involved to outsource their data on cloud servers for great convenience in data management and mining. However, data sets like health records in electronic documents usually contain sensitive information, which brings about privacy concerns if the documents are released or shared to partially un trusted third-parties in cloud. A practical and widely used technique for data privacy preservation is to encrypt data before outsourcing to the cloud servers, which however reduces the data utility and makes many traditional data analytic operators like keyword-based top-k document retrieval obsolete. In this paper, we investigate the multi-keyword top-k search problem for big data encryption against privacy breaches, and attempt to identify an efficient and secure solution to this problem. Specifically, for the privacy concern of query data, we construct a special tree-based index structure and design a random traversal algorithm, which makes even the same query to produce different visiting paths on the index, and can also maintain the accuracy of queries unchanged under stronger privacy. For improving the query efficiency, we propose a group multi-keyword top-k search scheme based on the idea of partition, where a group of tree-based indexes are constructed for all documents. Finally, we combine these methods together into an efficient and secure approach to address our proposed top-k similarity search. Extensive experimental results on real-life data sets

ISSN: 2366-1313

demonstrate that our proposed approach can significantly improve the capability of defending the privacy breaches, the scalability and the time efficiency of query processing over the state-of-the art methods.

1.INTRODUCTION

Cloud computing has emerged as a disruptive trend in both IT industries and research communities recently, its salient characteristics like high scalability and pay-asyou-go fashion have enabled cloud consumers to purchase the powerful computing resources as services according to their actual requirements, such that cloud users have no longer need to worry about the wasting on computing resources and the complexity on hardware platform management. Nowadays, more and more companies and individuals from a large number of big data applications have outsource their data and deploy their services into cloud servers for easy data management, efficient data mining and query processing tasks. But when the companies and individuals enjoy these advantages in cloud computing, they also need to take the privacy concern of the outsourced data into account. Because data sets in many applications often contain sensitive information like e-mails, electronic health records and financial transaction records, when the data owner outsourcing such sensitive data to the cloud servers which are considered to be partially trusted, the data can be easily accessed and analyzed by cloud service providers illegally. Since the analysis of these data sets may provide profound insights into a number of key areas in society (such as e-research, healthcare, medical and government services), thus data owners need effective, scalable and privacy-preserving services before releasing their data to the cloud.

Data encryption has been widely used for data privacy preservation in data sharing scenarios, it refers to mathematical calculation and algorithmic scheme that transform plaintext into cipher text, which is a non-readable form to unauthorized parties. A variety of data encryption models have been proposed and they are used to encrypt the data before outsourcing to the cloud servers. However, 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. The keyword-based search is such one widely used data operator in many database and information retrieval applications, and

ISSN: 2366-1313

its traditional processing methods cannot be directly applied to encrypted data. Therefore, how to process such queries over encrypted data and at the same time guarantee data privacy becomes a hot research topic. Fortunately, many methodologies based on searchable encryption have been studied. For example, deal with the single keyword search, and works support the multi-keyword boolean search. However, the single keyword search is not smart enough to support advanced queries and the boolean search 2 is unrealistic since it causes high communication cost. Therefore, more recent works like focus on the multikeyword ranked search, which is more practical in pay-asyou-go cloud paradigm. But 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. Motivated by this, in this paper, we focus on a special type of multi-keyword ranked search, namely the multi keyword top-k search, which has been a very popular database operator in many important applications, and only needs to return the k documents with the highest relevance scores.

For supporting multi-keyword search, we introduce the vector space model which represents documents and queries as vectors. In order to support top-k search, the relevance scores between documents and queries should be calculated, therefore, the TF×IDF (term frequency × inverse document frequency) model is introduced as a weighting rule to compute the relevance scores for ranking purposes. In addition, to improve the query efficiency for better user experiences, we propose a group multi-keyword topk search scheme (GMTS), which is based on partition and supports top-k similarity search over encrypted data. In this scheme, the data owner divides the keywords in the dictionary (suppose that the dictionary contains all the keywords that could be extracted from all documents) into multiple groups and establishes a searchable index for each group. On the other side, to better control the size of indexes, we adopt champion lists into our scheme, where the index of a keyword group only stores the top-ck documents of the corresponding keyword (the top-ck documents of a keyword represent the $c * k$ documents that have the highest relevance scores to this keyword, where c is a positive integer). Furthermore, we propose a random traversal algorithm (RTRA) to strengthen the data security, where the data owner builds a binary tree as searchable index and assigns a

random switch to each node, so the data user can assign a random key to each query.

Therefore, the data user can change the results and visiting paths of queries by using different keys, which maintains high accuracy of queries. Finally, we combine the GMTS and the RTRA together into an efficient and secure solution to our proposed problem. Our contributions can be summarized as follows: We first propose the random traversal algorithm which makes the cloud server randomly traverse on index and returns different results for the same query, and in the meantime, it maintains the accuracy of queries unchanged for higher security Based on the random traversal algorithm, we present one both efficient and secure searchable encryption scheme, which can support top-k similarity search over encrypted data. In this scheme, the data owner can control the level of query un link ability without sacrificing accuracy. Our experimental results show that our methods are more efficient than the state-of-the-art methods and can better protect data privacy. Especially, our proposed method has good scalability performance when dealing with large data sets.

1.1 PROBLEM STATEMENT

In order to protect the privacy of personal data, the data owners encrypt the documents before outsourcing them to the cloud. However, the encrypted data may make the traditional search mechanism based on plaintext keyword search obsolete. How to efficiently retrieve the encrypted data from the cloud has become a serious challenge.

1.2 DESCRIPTION

Inspired by the aforementioned problem, we present a fast and accurate multi-keyword ranked search scheme over encrypted cloud data, which supports the retrieval of top-k most relevant documents with high efficiency and accuracy. The FASE scheme utilizes FHOPE to encrypt the index and query vectors.

The FHOPE can support homomorphism addition, homomorphism multiplication, and order comparison over encrypted data. Therefore, the FHOPE can support the calculation of relevance score over encrypted data, and the relevance score is cipher text, which will not lose order. The cloud server can perform ranking operation on the relevance score without information leakage. And the dummy keywords are not added to the query vector and document vector, it retrieves the search results through the exact calculation of query vector and

ISSN: 2366-1313

document vector. We construct the document mark vector and the query mark vector, a large number of irrelevant documents are effectively filtered by matching the document mark vector and query mark vector, and the time cost for calculating the relevance score and ranking can be significantly reduced. When the cloud server performs the ranking operation, the search results are firstly ranked based on the keyword matching degree. If different documents have the same keyword matching degree, we rank the search results again based on relevance scores, so the ranking will be more accurate.

2.LITERATURE SURVEY

A. Single Keyword Searchable Encryption

Traditional single keyword searchable encryption schemes usually build an encrypted searchable index such that its content is hidden to the server unless it is given appropriate trapdoors generated via secret key(s). It is first studied by Song et al. in the symmetric key setting, and improvements and advanced security definitions are given in Goh , Chang et al. and Curtmola et al. Our early work solves secure ranked keyword search which utilizes keyword frequency to rank results instead of returning undifferentiated results. However, it only supports single keyword search. In the public key setting, Boneh et al. present the first searchable encryption construction, where anyone with public key can write to the data stored on server but only authorized users with private key can search. Public key solutions are usually very computationally expensive however. Furthermore, the keyword privacy could not be protected in the public key setting since server could encrypt any keyword with public key and then use the received trapdoor to evaluate this ciphertext.

B. Boolean Keyword Searchable Encryption

To enrich search functionalities, conjunctive keyword search over encrypted data have been proposed.These schemes incur large overhead caused by their fundamental primitives, such as computation cost by bilinear map, e.g. communication cost by secret sharing. As a more general search approach, predicate encryption schemes are recently proposed to support both conjunctive and disjunctive search. Conjunctive keyword search returns “all-or-nothing”, which means it only returns those documents in which all the keywords specified by the search query appear;

disjunctive keyword search returns undifferentiated results, which means it returns every document that contains a subset of the specific keywords, even only one keyword of interest. In short, none of existing Boolean keyword searchable encryption schemes support multiple keywords ranked search over encrypted cloud data while preserving privacy as we propose to explore in this paper. Note that, inner product queries in predicate encryption only predicates whether two vectors are orthogonal or not, i.e., the inner product value is concealed except when it equals zero. Without providing the capability to compare concealed inner products, predicate encryption is not qualified for performing ranked search. Furthermore, most of these schemes are built upon the expensive evaluation of pairing operations on elliptic curves.

3.SYSTEM DESIGN

3.1 System Architecture

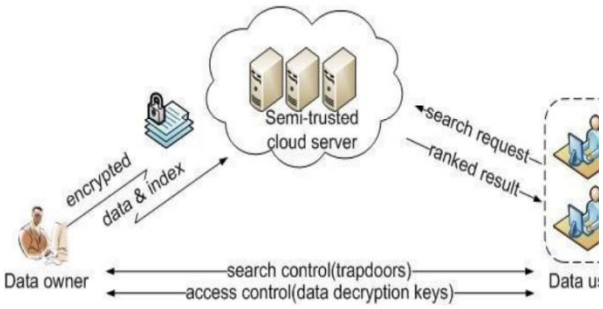

Fig 3.1: System Architecture

3.2 MODULES

: ❖ System Model

❖ Data User Authentication

❖ Illegal Search Detection

❖ Search over Multi-owner

3.3 Activity diagram:

➢ It is behavioral diagram which reveals the behavior of a system. it sketches the control flow from initiation point to a finish point showing the several decision paths that exist while the activity is being executed.

➢ This doesn't show any message flow from one activity to another, it is sometimes treated as the flowchart. Despite they look like a flowchart, they are not.

➢ In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system.

ISSN: 2366-1313

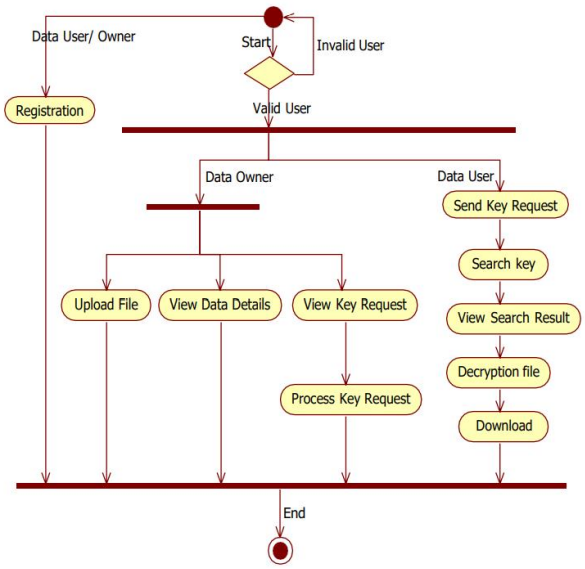

Fig.3.3 Activity diagram

4.OUTPUT SCREENS

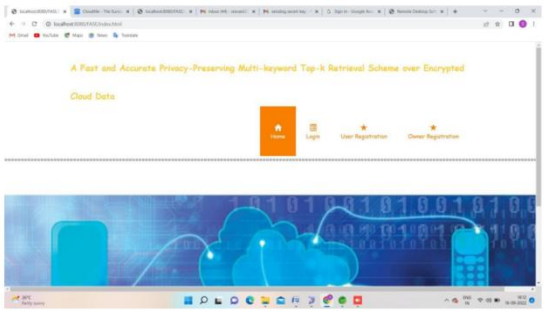

Fig 4.1 Home Page

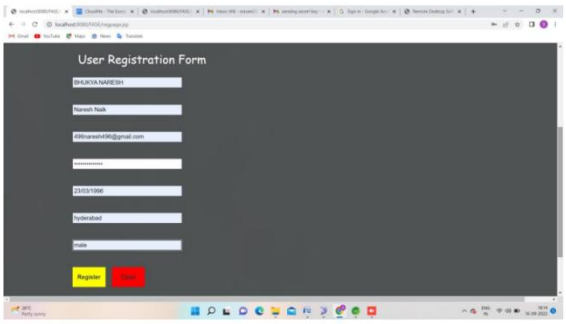

Fig 4.2 User Registration

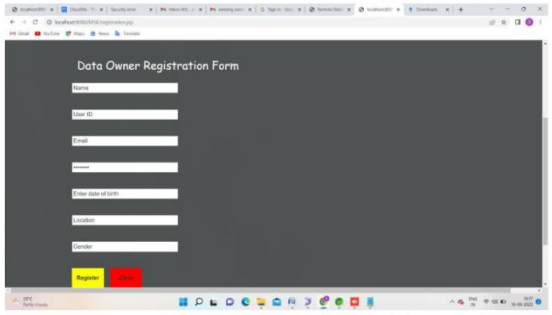

Fig 4.3 Data Owner Registration

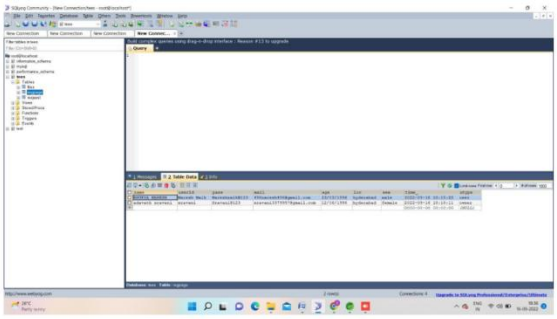

Fig 4.4 Registration details of data owner and data user store in database

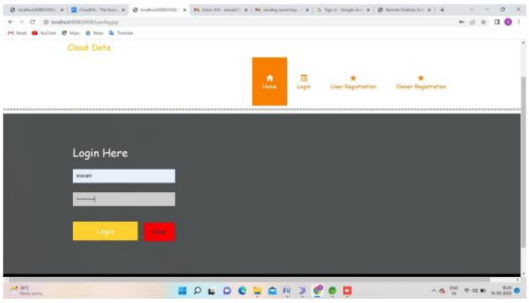

Fig 4.5 Data Owner Login

ISSN: 2366-1313

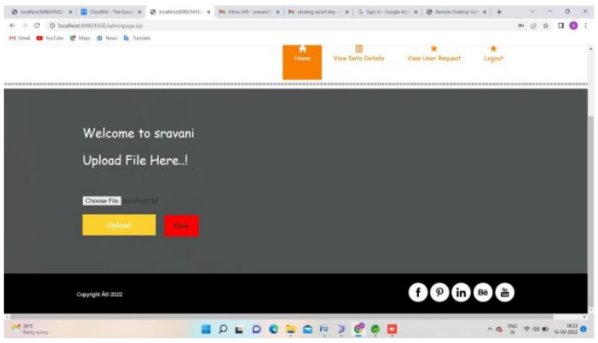

Fig 4.6 Data Owner File Upload

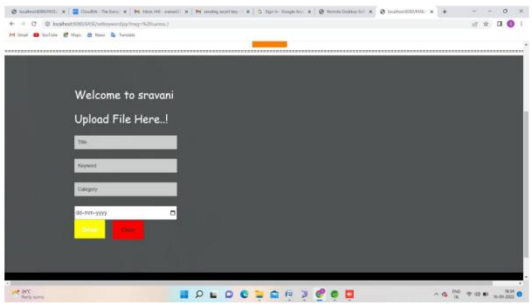

Fig 4.7 Data Owner File Key and Data Attaching

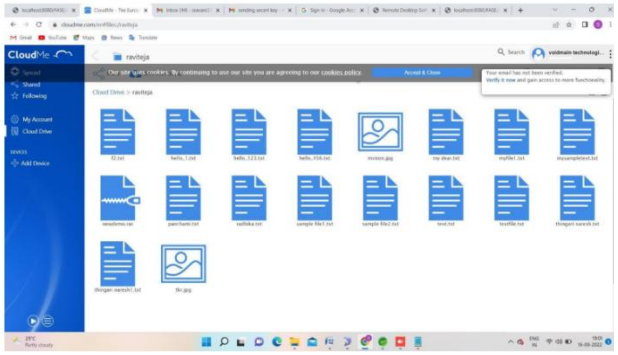

Fig 4.8 Data Store in Cloud

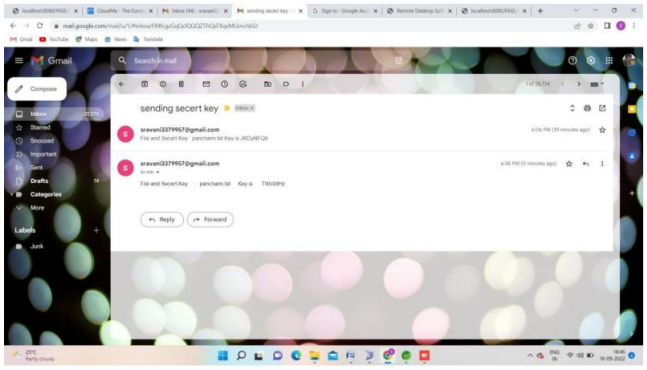

Fig 4.9 Data Owner Sent Key Through User Register Mail

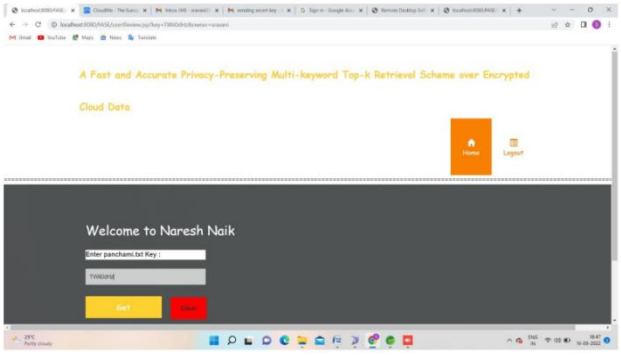

Fig 4.10 User Enter the Key in Given Field of Key

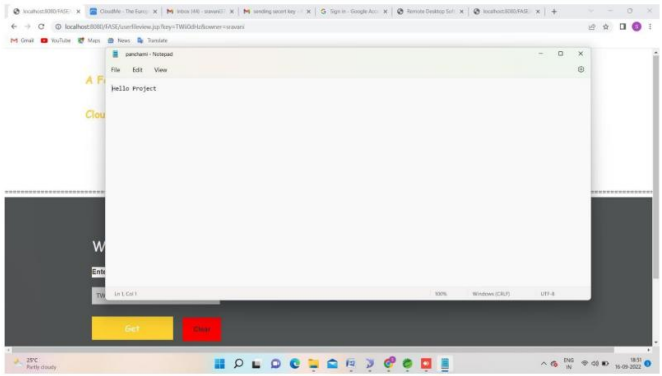

Fig 4. 11 Data User have Successfully Access File(Download)

5.CONCLUSION

In this paper, we focus on improving the efficiency and the security of multi-keyword topk similarity search over encrypted data. At first, we propose the random traversal algorithm which can achieve that for two identical queries with different keys, the cloud server traverses different paths on the index, and the data user receives different results but with the same high level of query accuracies in the mean time. Then, in order to improve the search efficiency, we design the group multi-keyword top-k search scheme, which divides the dictionary into multiple groups and only needs to store the top-ck documents of each word group when building index. Next, to protect the query unlinkability, we apply the random traversal algorithm to get the RGMTS, which can increase the difficulty of cloud servers to conduct linkage attacks on two identical queries, and we can also tune the value of E to make the level of query unlinkability flexible for data owners. Finally, the experimental results show that our methods are more efficient and more secure than the state-of-the-art methods.

6.FUTURE ENHANCEMENT

However, there are still some challenges in FASE scheme. As the documents stored at server may be deleted or modified and new documents may be added to the original data collection, a mechanism which supports dynamic operations is important. In the future work, we will try to design a dynamic searchable encryption scheme that supports dynamic operations, it could be a meaningful but difficult work.

7.REFERENCES

[1] S.S.M. Chow, Y.J. He, L.C.K. Hui, and S.-M. Yiu, “SPICE – Simple Privacy-Preserving IdentityManagement for Cloud Environment,” Proc. 10th Int’l Conf. Applied Cryptography and Network Security (ACNS), vol. 7341, pp. 526-543, 2012.

[2] L. Hardesty, Secure Computers Aren’t so Secure. MIT press, http://www.physorg.com/news176107396.html, 2009.

[3] C. Wang, S.S.M. Chow, Q. Wang, K. Ren, and W. Lou, “Privacy-Preserving Public Auditing for Secure Cloud Storage,” IEEE Trans. Computers, vol. 62, no. 2, pp. 362-375, Feb. 2013.

[4] B. Wang, S.S.M. Chow, M. Li, and H. Li, “Storing Shared Data on the Cloud via

ISSN: 2366-1313

SecurityMediator," Proc. IEEE 33rd Int'l Conf. Distributed Computing Systems (ICDCS), 2013.

[5] S.S.M. Chow, C.-K. Chu, X. Huang, J. Zhou, and R.H. Deng, "Dynamic Secure Cloud Storage with Provenance," Cryptography and Security, pp. 442-464, Springer, 2012.

[6] D. Boneh, C. Gentry, B. Lynn, and H. Shacham, "Aggregate and Verifiably Encrypted Signatures from Bilinear Maps," Proc. 22nd Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT '03), pp. 416-432, 2003.

[7] M.J. Atallah, M. Blanton, N. Fazio, and K.B. Frikken, "Dynamic and Efficient Key Management for Access Hierarchies," ACM Trans. Information and System Security, vol. 12, no. 3, pp. 18:1-18:43, 2009.