Efficient Privacy-Preserving Two-Round Multi-Keyword Top-k Similarity Search Over Cloud Data

M. Varshiny [1], M. Sindhukala [2], B. Shali Nayaki [3], I. Stephie Rachel [4]
[1, 2, 3, 4] Department of Computer Science and Engineering, University College of Engineering, Nagercoil, Tamil Nadu, India.

Abstract – A pool for sharing resources with high computing power and storage capacities that supports bigdata applications in various domains. With increasing popularity of cloud computing, the data owners are motivated to outsource their sensitive data to cloud servers for flexibility and reduced cost in data management. In these service, users are sending their personal data to a cloud-hosted server that stores the information for later access. Here, the cloud server is considered to be a partially untrusted third party so the privacy need to be concerned .because the datasets may contain sensitive information's. For data privacy preservation encryption technique is used that is the data is encrypted before outsourcing to the cloud servers. In this paper, proposed two-round multi-keyword cipher text retrieval TRMCR for top-k search problem for big data applications against privacy breaches, and attempt to identify an efficient and secure solution to this problem. Specifically, for the privacy concern of query data, it constructs a special block-based index structure and design a greedy depth first search algorithm, which makes the security stronger and the queries can easily be traversed. For improving the query efficiency, we use semantic query based content processing algorithm which gives result based on the search queries. Finally, we combine these methods together and attempts to provide a secure and efficient way to implement the proposed search. Various experimental results are demonstrated on our proposed search and it can improve privacy and time efficiency on query processing can be achieved.

Index Terms – Cloud computing, security enhanced, Encryption, two round multi keyword top-k search, greedy depth first search.

1. INTRODUCTION

Cloud computing , a pool for sharing resources that have been raised as an intentive resource in IT industries and research communities recently, its features like massive storage ,flexibility and self-service provisioning that is computing resources is used for any kind of work.Cloud computing is the delivery of computing services servers, storage, databases, networking, software, analytics and more over the Internet. With the exponential increase in data use that has accompanied society's transition into the digital 21st century, it is becoming more and more difficult for individuals and organizations to keep all of their vital information, programs, and systems up and running on in-house computer servers. Cloud computing provides solution to this problem. Pay-per-use allows users to pay for only the services and resources they use. The consumers can purchase powerful computing resources according to their needs. Cloud computing provides huge resources for the cloud users so that they can gain any types of services. They need not worry about computing resources and hardware platform management. Recently, all the users and employers outsource their data and services into cloud servers for easy data management, efficient data mining and query processing tasks. They enjoy the merits of the cloud, but the outsourced data's security need to be concerned. The outsourced data are said as datasets. The data owners outsourcing their data may contain sensitive information's like e-mails, electronic health records and financial transaction records and these sensitive information need to be secured. Because the cloud server is partially trusted, and the outsourced data can be easily modified or accessed by the cloud service providers illegally. So the released data need effective, scalable and privacy preserving services and these datasets may provide profound insights profound insights into a number of key areas in society. For data privacy preservation, data encryption techniques have been used. In this technique the original data is transformed into cipher text and it is in non-readable form which cannot be accessed by the unauthorized third parties that is it refers to the mathematical calculation and algorithmic scheme. Before outsourcing the data into the cloud servers, the data are encrypted. There have been proposed variety of encryption models. The traditional data processing methods that have been proposed for plaintext won't work well over encrypted data, however applying these encryption approaches may cause costs high in terms of data utility. The keyword-based search is such one widely used data operator in many database and information retrieval applications, and its traditional processing methods cannot be directly applied to encrypted data. Many methodologies have been proposed on searchable encryption. For example, dealing with single keyword search that supports the multi keyword search but these single keyword search is not enough for advanced queries and these costs high in communication cost since Boolean search is unrealistic. Simultaneously the high search efficiency and data security cannot be achieved especially for bigdata applications and it poses great scalability and efficiency schemes.

The proposed system can be summarized as:

- We first propose the greedy depth first algorithm which makes the cloud server to randomly traverse the on block and provide different results based on the relevant scoring.
- Based on the greedy depth first search algorithm we attempt to provide efficient and secure searchable encryption over encrypted data.

2. RELATED WORK

Searching over encrypted data is considered to be a tedious task, concerning with the cloud computing. In this section, search over encrypted data are analysed and reviewed. [3]New protocols that provides secure search among multiple owners data and additive order to rank the search result.[4] So it's more efficient on large data on keyword sets. But the problem here is it does not support secure fuzzy keyword search in a multi-owner data. [6]An efficient multi keyword fuzzy ranked search scheme in which it is tolerance to single and multiple misspelling of letters in keywords. [5]Here exact search and single keyword fuzzy search are combine to provide efficient and accurate query processing. But here some keywords like anagrams cannot be distinguished. By conducting various experimental results dynamic data operation is supported in cloud environment. By analysing these results be attempt to provide approximate search result on query processing

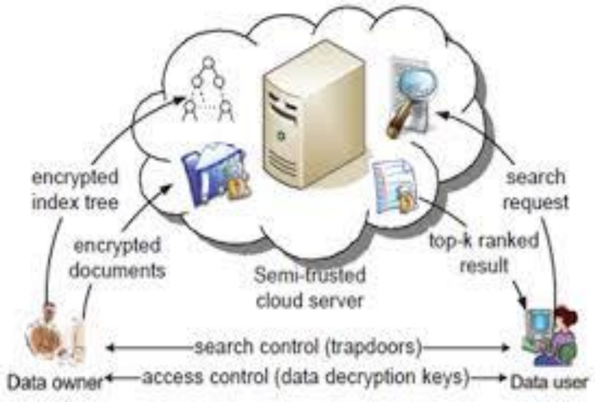

Figure 1: searching model for outsourced encrypted data

3. PROPOSED SYSTEM

We propose a secure search over encrypted cloud data. Here the data owner can upload their files into the cloud server which is encrypted twice for security purposes. The data user can search for the files in the cloud server but they cannot access the files or view it. Only authorised users can view it. The data owner and data user details are verified by the cloud service provider for authentication purposes. In our proposed system search efficiency is achieved. The search results are provided based on the highest relevance scores of the searched files. Here we use Greedy depth first algorithm that searches all the files to provide a better search over encrypted data. The semantic query based content aware processing technique provides query based search over encrypted data.

3.1 System model

According to the fig 1 the system design in this paper consists of three parts: data owner, data user and cloud server. Here the data owner can act as both data user and data owner .the data owner uploads files to the cloud server in a collection of files F. These files are stored in cloud server in encrypted format as a encrypted file collection E. Here the service providers is considered to be a untrusted third party. The data stored in cloud server may contain sensitive information, so it is stored in the encrypted format. The data owner constructs a searchable index block Br locally in which data users can search for the files. Finally the files collection and the encrypted searchable index are stored in the cloud server. [7]When the data user search for the files in cloud server, the cloud server calculates the relevant scoring and submits the results for the search query. If the necessary files are chosen it cannot be viewed by the data users. [9]The data users need to request to the owners of the files, so that request are send to the data owners. After that the decryption keys are send to the authorised data user. [2]The data owners can view the files uploaded by him/her. Here we assume that the data owners and the data users are considered to be a trusted parties. [8]We treat the cloud service providers as the untrusted third parties. But the servers are honest as it performs the algorithms and programs correctly. [1]The cloud service providers can analyse and easily access the data stored in the cloud servers. Here the model corresponds to the cipher text-only attack since the server only knows the encrypted files collection E, encrypted searchable block Br. Compared to this the cloud server also knows various other knowledge such as the relevant scoring of the files stored. So the proposed model provides efficient search efficiency. The below table represents the notations used in this paper.

Table 1

Notations	Description
F	The plain text file collection denoted as F=(F1,F2,……Fm) Fi is denoted as a file of F
E	The encrypted file collection
B	The unencrypted form of searchable block
Br	The encrypted of searchable block B
D	The dictionary which contains keywords in the file collection F

3.2 Two round multi keyword top-k search

Let F be a plain text file collection in the cloud servers that were outsourced by the data owners.Then let Fi be a file from the file collection in F.D is the dictionary and score (Q, Fi) is the relevance score between query Q and file Fi. The two round multi keyword top-k search is used find the files with the highest relevance scoring.

For example, file collection F has 3 files as shown in fig 2 in which each files is represented as vectors. These vectors stores the TF values.

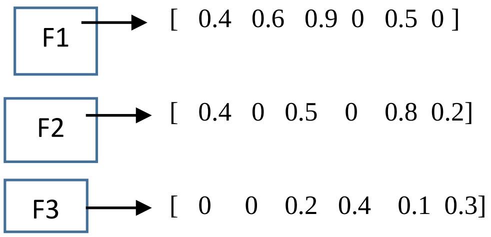

Figure 2: Each file is represented as a vector in file collection.

By our assumption, it is clear that here the files F1 and F2 have the highest score.

4. DESIGN GOALS

Our goal is to provide a efficient search over encrypted data and to achieve the search efficiency and also to provide security measures 'Coordinate matching' with top-k similarity search over encrypted data.

Security

The block and queries should be secured from the untrusted third parties. The information contained in the plain text files and the encrypted searchable block should be secured.

Search Efficiency

Our proposed system should provide better search efficiency compared to other state-of-art-methods in search and query processing and it will be more efficient and effective. information of the search results should be hidden from the cloud servers.

5. GREEDY DEPTH FIRST TRAVERSAL OR DFS FOR A GRAPH

The Greedy depth first algorithm is a recursive algorithm that uses the idea of backtracking. It traverse into all nodes until it reaches the end and then the backtracking is done. Here, the word backtrack means that when you are moving forward and there are no more nodes along the current path, you move backwards on the same path to find nodes to traverse. All the nodes will be visited on the current path till all the unvisited nodes have been traversed after which the next path will be selected. This recursive nature of DFS can be implemented as blocks. The basic idea is as follows: Pick a starting node and also its adjacent nodes into a block. Delete a node from block to select the next node to visit and insert all its adjacent nodes into a block.

Repeat this process until all the nodes are traversed. However, ensure that the nodes that are visited are marked. This will prevent you from visiting the same node more than once. If you do not mark the nodes that are visited and you visit the same node more than once, you may end up in an infinite loop. The figure 3 depicts that all the files are stored as nodes such as f1, f2, f3, f4, f5, f6, f7, f8, f9 f10. For example the file to be traversed is f5. Initially all the files are traversed and the searched file is found.

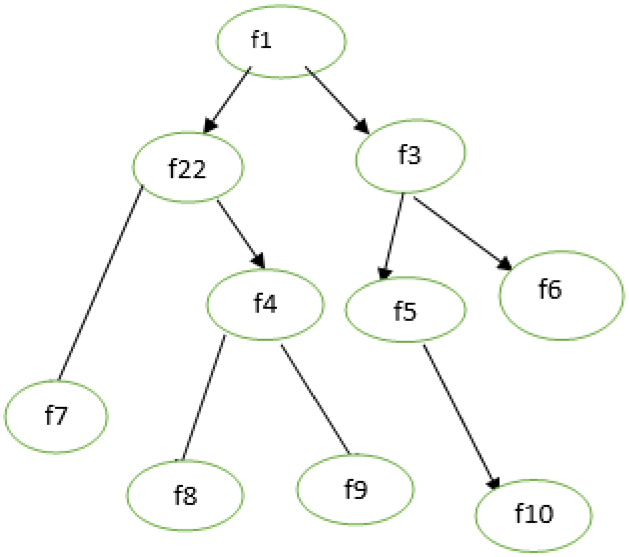

Figure 3: Files that are traversed using greedy depth first algorithm

6. ARCHITECTURE

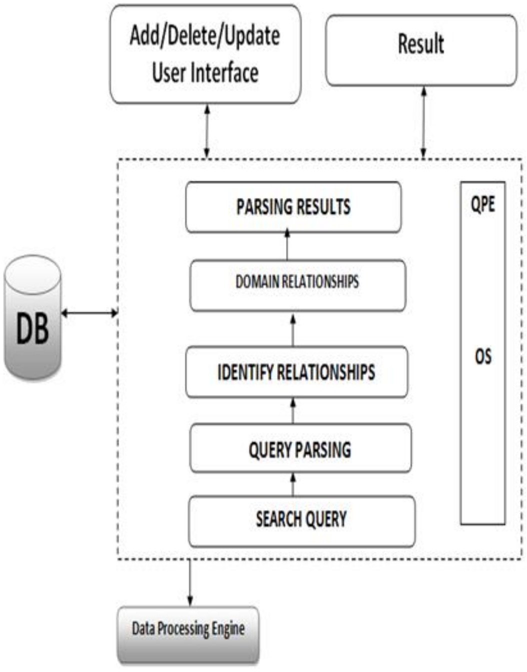

Figure 4: Architecture of the proposed system

The architecture consists of three parts: Database, Data Processing Engine and User Interface. Here the database stores all the files and document that are uploaded by owners. In data processing there are five units. Search query, query parsing, identify relationship, domain relationship, parsing results.

In search query, the query to be searched is given for parsing the search query is parsed by the query parser. After query parsing, it identify the relationships of the given query and the domain relationship is found and the result is parsed. Here the domain relationship refers to the relationship between the queries. The OS and QPE refers to the operating system and query processing engine respectively. The query processing engine controls all the units of the data processing engine.

For example if the search query is "cloud computing" here the query is found and the domain relation are found finally the result is given based on the highest relevance scoring.

- In multi-keyword ranked search we use Vector Space Model (VSM) to build the index, where each file is expressed as a vector in which each dimension values is the Term Frequency (TF).
- An index tree has been constructed by using the index vector.
- The related file document can be found by traversing the tree.

7. ALGORITHM

Query Processing and Data Processing Technique (semantic query based content aware processing algorithm)

```
for each block $B_r$ of$r$ do begin
  for each block $B_s$ of $s$ do begin
    for each tuple $t_r$in $B_r$ do begin
      for each tuple $t_s$in $B_s$do begin

          Check if $(t_r,t_s)$ satisfy the join condition
if they do, add $t_r$• $t_s$to the result.

end

end
   end
end
```

8. CONCLUSION AND FUTURE WORK

We attempt to improve the security and efficiency of two round multi keyword top-k similarity search over encrypted data. At first we propose the greedy depth first algorithm through which we can achieve search efficiency by traversing different paths on the block. The results are provided based on the relevant scores to the data users. In order to improve the search efficiency we design semantic query based content aware processing technique, which search the query based on the provided content. For security purposes the data outsource to the cloud server and encrypted twice. Only authorised users can access those encrypted files. Finally we combine all these methods an provide a more efficient and secure search over encrypted cloud data through many experimental results we proposed that our methods are more secure and efficient than the sate-of-the-art-methods. The future work will focus on the extension of large scale system to hadoop based systems. To identify the drawback of the system and will improve the data response time. It further improved by using semantic-aware namespace to provide dynamic and adaptive namespace management for ultra large storage systems in hybrid systems.

REFERENCES

[1] Zhangjie Fu, Xinle Wu, Chaowen Guan, Xingming Sun, Kui Ren, "Toward Efficient Multi-Keyword Fuzzy Search Over Encrypted Outsourced Data With Accuracy Improvement" IEEE Transactions On Information Forensics And Security, Vol. 11, No. 12, December 2016.

[2] Hongwei Li, Yi Yang,Tom H. Luan, Xiaohui Liang, Liang Zhou, Xuemin (Sherman) Shen, "Enabling Fine-Grained Multi-Keyword Search Supporting Classified Sub-Dictionaries over Encrypted Cloud Data", IEEE Transactions On Dependable And Secure Computing, Vo. 13, No. 3, May/June 2016

[3] Wei Zhang, Yaping Lin, Sheng Xiao, JieWu, Siwang Zhou, "Privacy Preserving Ranked Multi-Keyword Search for Multiple Data Owners in Cloud Computing", IEEE Transactions On Computers, Vol. 65, No. 5, May 2016.

[4] Hui Cui, Zhiguo Wan, Robert H. Deng, Guilin Wang, Yingjiu Li, "Efficient and Expressive Keyword Search Over Encrypted Data in Cloud", IEEE Transactions on Dependable and Secure Computing Journal Of , Vol. , No., 2016.

[5] Chi Chen, Xiaojie Zhu, Peisong Shen, Jiankun Hu, Song Guo, Zahir Tari, Albert Y. Zomaya,.

[6] Zhangjie Fu, Kui Ren, Jiangang Shu, Xingming Sun, Fengxiao Huang", Enabling Personalized Search over Encrypted Outsourced Data with Efficiency Improvement," IEEE Transactions On Parallel And Distributed Systems, Vol. 27, No. 9, September 2016.

[7] Zhihua Xia, Xinhui Wang, XingmingSun,Qian Wang, Member, "A Secure and Dynamic Multi-Keyword Ranked Search Scheme over Encrypted Cloud Data", IEEE Transactions On Parallel And Distributed Systems, Vol. 27, No. 2, February 2016.

[8] Jingbo Yan, Yuqing Zhang, Xuefeng Liu, "Secure Multi-keyword Search Supporting Dynamic Update and Ranked Retrieval", Services and applications, China Communications, 2016.

[9] Chia-Mu Yu, Chi-Yuan Chen, and Han-Chieh Chao, "Privacy-Preserving Multi-keyword Similarity Search Over Outsourced Cloud Data", 1932-8184 © 2015 IEEE Systems Journal.

[10] Wenhai Sun, Bing Wang, Ning Cao, Ming Li, Wenjing Lou, Y. Thomas Hou, Hui Li, "Verifiable Privacy-Preserving Multi-Keyword Text Search in the Cloud Supporting Similarity-Based Ranking", IEEE Transactions On Parallel And Distributed Systems, Vol. 25, No. 11, November 2014.