ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

MIR-tree Indexed Multi-Keyword Ranked Search Scheme over encrypted Cloud Data

Sherin T Thayyil[#1], Kiran G Kumar [*2]

[#]*Computer Science And Engineering*
Gurudeva Institute of Science And Technology (GISAT) Kottayam, Kerala, India
sherinnasa@gmail.com
[*]*Dept.of CSE, GISAT*
kirangokulam@gmail.com

***Abstract*— cloud computing is the most popular technology among the internet and computers. Data owners may outsource their sensitive data to the cloud for flexibility and reduced cost in data management. Privacy is a big concern for outsourcing data to the cloud. The data owners typically encrypt documents before outsourcing for privacy-preserving. As the volume of data is increasing at a dramatic rate, it is essential to develop an efficient and reliable cipher text search techniques, so that data owners can easily access and update cloud data. In this paper, we propose privacy enabled multi-keyword ranked search scheme over encrypted data in cloud along with data integrity using a new authenticated data structure MIR-tree. The MIR tree based index with including the combination of widely used vector space model and TF×IDF model in the index construction and query generation. The method use inverted file index for storing word-digest, which provides efficient and fast relevance between the query and cloud data. Because of tree based index, the scheme achieves optimal search efficiency and reduces communication overhead for verifying the search results. The analysis shows security and efficiency of proposed method.**

Key words: Cloud computing, MIR-tree, multi-Keyword ranked search, data integrity, privacy-preserving, trapdoor.

I. INTRODUCTION

Cloud computing environment provides huge resources of computing, storage and ease of accessing data in a most secured way with great efficiency and less operating cost . So, enterprises and data owners usually choose to outsource their data to cloud in order to avoid data management at locally. As a result, more people are moving their data to the cloud. Despite the benefits, privacy is a big concern for cloud storage. Although cloud service providers provide strong security mechanism, but there are always chances of leakage of confidential data (for example personal includes emails, tread secrets etc.) or intruders may access users' data without authorization. Privacy preserving and secure storage are two main concerns about outsourcing data on cloud. All papers will be reviewed prior to the symposium and must meet technical standards. Awards for excellent papers will be presented during the symposium period.

One of the popular ways to ensure privacy preserving is to encrypt the data before outsourcing. Sometimes, data owners may share their data with authorized users, and users retrieve the data files from the cloud based on the keyword search techniques. But, keyword search on cipher text is a challenging task because of limited operations on cipher text. Also, it is preferred to get most relevant files which user need, so searched files should be ranked in order of text relevance. Only top relevant files are sent back to user for fast and accurate results.

In this paper, we propose a solution to address the problems of existing systems, by considering a secure MIR-tree based multi-keyword search over the encrypted data and search result verification. In our model, we create a MIR-tree-based index, which provide authenticate text relevance by using the word digests. This construction uses word digests including term frequency (TF) and inverse document frequency (IDF) ,and vector space model is used for multi-keyword search and query generation Then, Greedy Depth-first Search (GDFS) algorithm is used on the MIR index tree for fast search and construct an authentication set for search result verification.

II.RELATED WORK

Recently several approaches are defined to protect the sensitive data which stored on the cloud database. The existing approaches mainly focus on the encrypted cloud data. It's clear that it's difficult to do the search in the encrypted data. The existing schemes build an index for each file and then encrypt both the file and the index and stored to the cloud server. To retrieve the files stored on the cloud server the user creates a trapdoor and search the specified files based on the trapdoor generated and the most relevant files are retrieved based on the ranking of the files.

In [1] Swaminathan et.al proposed a model for privacy of encrypted data based on indexing security. Cao et.al [2] discussed a search scheme to get back relevant files by using
techniques of coordinate matching. Xia et.al [3] proposed a scheme which used a tree based index generation along with vector space model. But these

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

models lack data integrity, so some times the results can be distorted or even chance to get the wrong results.

To verify the search results there are other schemes also proposed in which Syam et.al [4] proposes a verifiable secure search scheme using two different kinds of keys, one for encryption and another for the decryption of the specified text. Chen et.al [5] proposes a scheme based on the hash tree. But these schemes take much communication and computation overhead for verifying search results.

III.PROBLEM SETTING

A. *Notations and Preliminaries*

- D - The collection of n documents containing,
  D = $d_1, d_2, \ldots..d_n$.
- n - The total number of documents in D.
- W - The dictionary containing all keywords used in the collection of all the documents D, denoted as
  W = $w_1, w_2, \ldots\ldots w_m$.
- m - The total number of keywords in W.
- Wq - The keywords in the query , a subset of W.*Length*
- The maximum length of your paper including figures and tables should not exceed four pages.
- Q - The query vector for the keyword set Wq.
- C - The encrypted form of D, containing C = c1,c2,.....,cn
- TD - The encrypted form of Q, namely search request from the client to server.
- T - The unencrypted form of MIR-based index tree.
- I - The encrypted form of the index tree.
- k - The number of top results wants to retrieve from search scheme.
- RS - Top-k results
- h() - A secure hash function.
- h(d1) - The hash value of document d1.
- h(h1|h2) - The concatenation of two hash values h1 and h2.
- hw(e) - The word digest of node e for word w.

B. *Vector Space Model*

It is one of the popular methods to measure text relevance on plaintext information retrieval. It provides accurate relevance results by using TF×IDF model, where TF means occurrence count of a keyword within a document i.e. it gives the number of times a keyword is present in a document and IDF is obtained by dividing the total number of documents by the number of documents containing that keyword.

C. *Inverted index and Word Digest*

Inverted index [13] is an efficient and popular index for keyword search, indexing the text descriptions of objects (documents). The word digests are stored in the inverted file.

D. *MIR Tree*

MIR-tree is used to authenticate the text relevance. In the MIR-tree [12], [13], each leaf node is the hash value of a document from the collection D.

Hence leaf nodes are computed by hashing the documents like, $h_1 = h(d1), h_2 = h(d2)$ and so on. To calculate the non-leaf nodes, we can take concatenation of two child nodes and assign their resultant hash value to their parent node,

For example: $h_{1-2} = h(h1|h2)$ and so on. Similarly complete tree is constructed and root node value is the concatenation of

hash values of its child nodes i.e. $h_{root} = h(h1-4|h5-8)$ as shown in Figure 1.

D. *Score Calculation*

The similarity score between a node and the query is calculated by using a rank function:

- ud,t : The TF of the keyword t within document d i.e. weight defined in word digest.
- ut : the number of documents containing the keyword t among the total number of documents n.
- vd,t: the TF weight for ud,t.,$vd,t = (1 + ln(ud,t))$

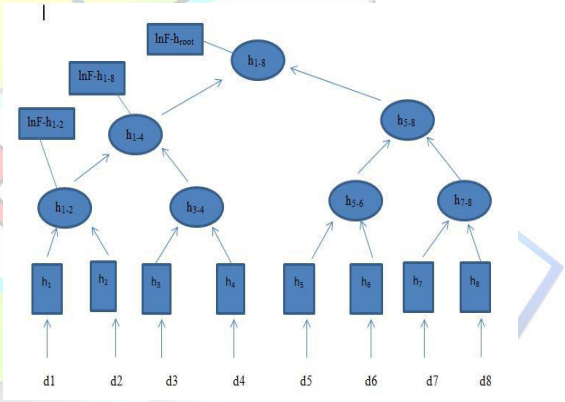

Fig 1: MIR Tree with 8 leaf nodes

- vQ,t : the IDF weight for the query vector $Q, vQ,t = (1 + n/ut)$

$$Rank(WDe,Q) = tQ\ vd,t.vQ,t \qquad (1)$$

E. *System Model*

The proposed system includes Data Owner (DO) Data user (DU) or client and Service Provider(SP) or the cloud server. Data owner has the collection of documents D that he needs to outsource to the cloud server. Data owner first builds an index I from D and then encrypts both D and C. Then, it outsources the encrypted C together with I to the cloud server also distributes secret keys and information about tree construction to the desired data users through a secret channel.

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

Data users have secret keys to access the documents of DO. They generate a Trapdoor T with some query keywords to perform search over encrypted data. After getting results from SP, the user verifies the search results, if that are authenticated, user decrypts the documents with secret keys.

Cloud server is the storage platform, it stores the encrypted Documents C and index I. After receiving a trapdoor, it starts search within index tree I and return the matched encrypted documents, which are ranked by our ranking method.

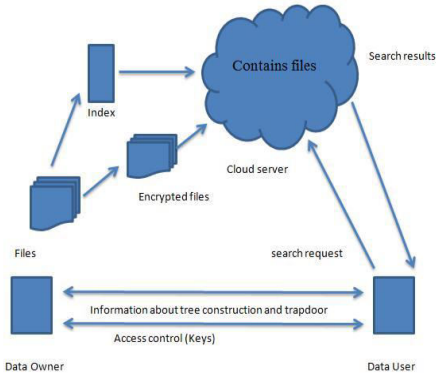

Fig 2: System Architectural model

IV. PROPOSED SOLUTION

The system propose a new approach based on MIR-tree for secure storing of sensitive data on cloud, and provide efficient search scheme along with integrity and authentication of search results, which make outsourcing of data a reliable system. One of the popular and efficient methods to construct authenticated data structure is by using Merkle hash tree (MHT). It is widely used for verifying the query results in a most appropriate way. In our scheme, instead of using MHT, we have used Merkle-IR-tree (MIR-tree) [13] to optimize the DO and SP communication overhead and computation overhead for the user for verifying search results. The proposed solution consists of two phases which are, the initialization phase and the retrieval phase Author details must not show any professional title (e.g., Managing Director), academic title (e.g., Dr.), or membership of any professional organization (e.g., Senior Member IEEE).

A. *Initialization Phase*

In initialization phase, DO pre-process the data,that include four algorithms which are as follows:

1. KeyGen():Secret Key k is produced by DO along with randomly generated two invertible mtrices M1 and M2, and a secret key of n bits.

2. BuildIndex(D,T,S,K): data Owner constructs a secure tree index from the collection of documents. The data owner computes the word digests $WD_{e,w}$ for each non-leaf node. It stores the TF values for keyword w.

   Then,it generates two random vectors(WD_e,WD_e) for non-leaf node e with word digest WDe. According to S, the splitting procedure is applied to WDe.

3. Enc(T,d,S,K) : DO encrypt it with a secure searchable encryption algorithm into index I,and document collection D to C.

B. *Retrieval phase*

Retrieve the documents stored in cloud based on user query.

1. TrapdoorGen(Q,SK): From the keywords in dictionary W,Wq are taken as a query vector Q with m number of keywords. For each w_i in Wq,Q [i] stores the IDF value of w_i, else Q[i] set to 0. Similarly, Q is split into two random vectors Qand Q. Finally, we have encrypted form of Q in form of trapdoor TD.

2. Search(): SP searches for the relevant documents to TD over encrypted tree.

3. Ranking(RS,K):Client computes the relevant score of objects using equation 1,ranks them and obtain their result.

4. Verifying search results: After ranking the returned results from the SP, the user verifies the search results.

5. Dec(RS,SK): If the search results are authenticated then, the data user utilizes the secret key SK to decrypt the result obtained.

V.CONCLUSION AND FUTURE WORK

The System proposes a new approach using MIR Tree. The MIR tree indexed architecture provides more privacy concerns and it limits the time consumption. The system ensures correctness and the completeness of the search results by the ranking method. The MIR tree based index is developed to make a secure and flexible index structure which supports efficient search or traversal with the help of the query trapdoor.

There are still many challenges to overcome so that the efficiency and security of our scheme are upgraded. In future work, we would like to implement dynamic operations on the collection of data, insertion, deletion and update.

REFERENCES

[1] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, Privacy- Preserving Multi-keyword Ranked Search over Encrypte D cloud data, Proc. *IEEE INFOCOM, 2011.*

ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

[2] A. Swaminathan, Y. Mao, G.-M. Su, H. Gou, A.L. Varna, S. He, M. Wu, and D.W. Oard, Confidentiality-Preserving Rank-Ordered Search Proc. Workshop Storage Security and Survivability, 2007.

[3] Xia, Zhihua, Xinhui Wang, Xingming Sun, and Qian Wang. ”A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data.” *IEEE Transactions on Parallel and Distributed Systems 27.*

[4] Chen, Chi, Xiaojie Zhu, Peisong Shen, Jiankun Hu, Song Guo, Zahir Tari, and Albert Y. Zomaya. ”An efficient privacy-preserving ranked keyword search method.” IEEE Transactions on Parallel and Distributed Systems 27, no. 4 (2016): 951-963.

[5] Pasupuleti Syam Kumar, Subramanian Ramalingam, and Rajkumar Buyya. ”An efficient and secure privacy-preserving approach for outsourced data of resource constrained mobile devices in cloud computing.” Journal of Network and Computer Applications 64 (2016): 12-22.