The Multi-keyword Synonym Search for Encrypted Cloud Data Using Clustering Method

Monika Rani H G[1], **Varshini Vidyadhar**[2]

M. Tech Scholar, CS&E, Acharya Institute of Technology, Bengaluru, India[1]

Asst. Professor, Dept. of CSE, Acharya Institute of Technology, Bengaluru, India[2]

Abstract: In the era of cloud computing, the large data owners are willing to store their precious data on cloud instead of their own server, to reduce maintenance burden and for low cost expenditure. This trend introduces the problem of data privacy for storing the sensitive data on cloud. Therefore, the confidential data has to be encrypted before outsourcing to cloud services. It is a challenging task to search an encrypted data because the relationship between plain text and encrypted text are concealed. We need an efficient search method for online retrieval of required data from the encrypted documents repository on cloud. We propose an efficient and secure search with hierarchical clustering method to improve the search efficiency and propose synonym search along with this method to increase the relevance of retrieved documents. This method ensures the Multi-keyword synonym search, relevance between documents, checks the ranking of documents and solves the problem of maintaining the relationship between plain text and encrypted text. It is proved with the help of practical analysis.

Keywords: Cloud Computing, Hierarchical clustering method, Multi-keyword synonym search.

I. INTRODUCTION

Due to the exponential growth of data, the data owners tend to store their data into the cloud to get rid of data storage and maintenance overhead. This trend makes almost all the data to be centralised in Cloud. Since we cannot trust the third party domain storage completely, there is a high risk of our outsourced data privacy and security. In Cloud storage scenario, there will be a third party storage so there will be chance of data insecurity and loss of privacy.

In order to prevent the data leakage, the common method used is encryption. Therefore, the data has to be encrypted before sending it to cloud. Searching in large volume of encrypted data is a challenging task. We need an efficient cipher text search method, which will give accuracy and preserving privacy of data.

This will lead to use an efficient index search method [1] like clustering hierarchical index and we introduces a vector space model, considering every documents as a vector in two dimensional space and the similar documents are grouped to form a cluster.

The reminder of this paper is organized as follows. In section 2, we describe the relative work; Section 3 provides the detailed description of our proposed method. Section 4 presents the experimental results. Section 5 gives the summary of the proposal.

II. RELATIVE WORK

In present years, searchable encryption, which provides text search function based on encrypted data, has been widely studied [7-8], especially in security definition, Formalizations and efficiency improvement.

The proposed method is compared with existing methodologies and has the advantage in maintaining the relationship between documents.

A. Searchable Encryption

Cryptography provides us the orthodox encryption search methods. Among those works, most are focused on efficiency improvements and security definition formalizations. Song et al proposed the first implementation of searchable encryption, where in every word in the document is individually encrypted using two-layered encryption construction [5].

Goh construct the indexes for the data files using bloom filters. A bloom filter having unique words trapdoors is built for each file and kept on server. User creates individual request for search by computing the trapdoor for the word and send to server, when searching for a word. Then server does process the request by checking the bloom filter containing the trapdoor of the query and corresponding file identifiers are returned. Several approaches are there for single word encryption, but none of them are suitable for cloud computing.

B. Fuzzy keyword search

In this scheme [2], the author investigates and resolves the problem of effective fuzzy keyword search on encrypted data stored over cloud repository while preserving keyword privacy [10-11]. This search method improves system usability by sending back the exact matched result, when user input query search matches absolutely with the previously defined keywords or else it will return the similar documents based on the semantic relevance of the

DOI 10.17148/IJARCCE.2016.5858

keyword requested by end users. This scheme uses edit distance method to measure keyword similarity.

This search method gives the output results based on below constraints: 1) When the user has input search query exactly matching the pre-stored keywords, the server produces the search output in terms of the files containing those keywords. 2) If there are any mismatches exists in types or format in the searching query words, the similar possible output based on pre-specified semantics will be presented by server to the clients.

C. MTS Scheme

In this paper [4], they propose a privacy-preserving multi-keyword text search (MTS) method with similarity-based ranking to address the problem of efficient search in encrypted cloud data. They use vector space model to ensure the grouping of similar documents that will make easy to access the relevant files in faster manner and use term frequency measure to develop the search keyword indexes. To facilitate the search results ranking accuracy, they will use cosine similarity measure. It is one of the better methods to check similarity among the files.

The built search indexes are formulated in tree based structure to enhance the search efficiency, which gives faster access to the documents. They adapted different multi-dimensional algorithm to increase the search efficiency than the linear search. They will satisfy the privacy factor by introducing several threat models.

D. Multi-keyword Ranked Search method

This method [3] [9] defines and solves the problem of multi-keyword ranked search over encrypted cloud data, and establish a variety of privacy requirements. Among various multi-keyword semantics, we choose the efficient similarity measure of “coordinate matching”, i.e., as many matches as 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, a basic idea of MRSE using secure inner product computation is proposed.

Then we give two improved MRSE schemes to achieve various stringent privacy requirements in two different threat models. Thorough analysis investigating privacy and efficiency guarantees of proposed schemes is given and experiments on the real-world dataset show our proposed schemes introduce low overhead on both computation and communication. This leads to further gap for supporting other multi-keyword semantics over encrypted data and checking the integrity of the rank order in the search result.

III. PROPOSED METHOD

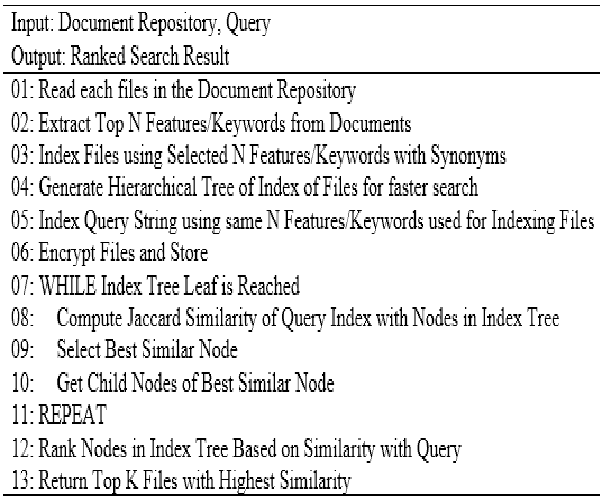
The method proposed will have vectors in two dimensional space, built from each documents in the repository. Extract keywords from the documents before encrypting and generate the synonyms for each keyword. This will form hierarchical cluster of indexes that increases the search efficiency. Cluster is a group of similar documents. Documents again form subgroups and this will continue till each document form a separate group. This forms the hierarchy of index to search the set of required documents.

The synonym search gives the relevance between the documents and clustering index method by greatly reducing the searching time as it will search only the sub group which is matched to the query keyword.

When the query is entered by the user, it will start searching from the root of the hierarchy, till the relevant subgroup matched. Then a rank the documents using Jaccard coefficient measure and gives the top ranked documents as a result for the users.

Fig.1 explains the steps involved in this method of implementation

All the documents and indexes are encrypted using AES algorithm



```
Input: Document Repository, Query
Output: Ranked Search Result
Read each files in the Document Repository
Extract Top N Features/Keywords from Documents
Index Files using Selected N Features/Keywords with Synonyms
Generate Hierarchical Tree of Index of Files for faster search
Index Query String using same N Features/Keywords used for Indexing Files
Encrypt Files and Store
WHILE Index Tree Leaf is Reached
    Compute Jaccard Similarity of Query Index with Nodes in Index Tree
    Select Best Similar Node
    Get Child Nodes of Best Similar Node
REPEAT
Rank Nodes in Index Tree Based on Similarity with Query
Return Top K Files with Highest Similarity
```

Fig.1: Algorithm

A. System Architecture

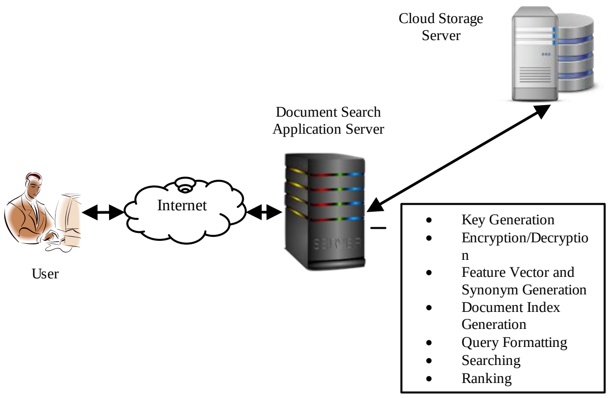


Fig. 2: System Architecture

This system architecture as shown in the Fig.2 explains the working components namely, User, Application Server and Cloud Storage. Whenever user enters the query keyword, it will be transferred over the internet and the application server process the user request by matching the index for the query keyword. It fetches the matched encrypted document from the cloud storage, decrypt it and send it back to the user. The user can specify the number of documents required in search results. The top k ranked documents will be fetched by the server and give it to the user. This will include number of steps such as Key Generation, Encryption/Decryption, Feature Vector and Synonym Generation, Document Index Generation, Query Formatting, Searching and Ranking.

IV. EXPERIMENTAL RESULTS

In this section, we present the experimental evaluation of the proposed technique. A set of 100 documents are taken for experimental evaluation. The results are calculated based on the number of keywords, Similarity measure, and number of keywords, document index. The comparison for with synonym search and without synonym search.
The search relevance is shown in the Fig 3. It can be seen that the document similarity is more with synonym search when compare with normal keyword search as the relevance increases with considering the synonyms along with the number of keywords.

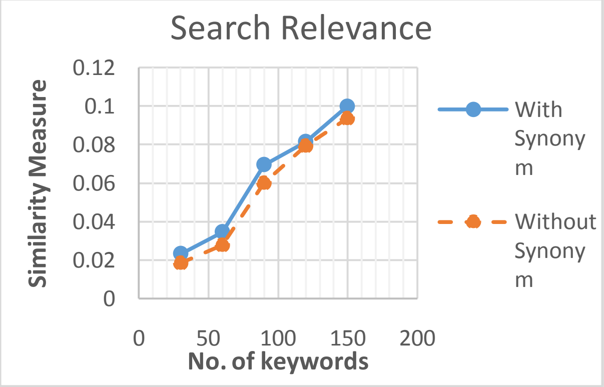


Fig.3 Search Relevance

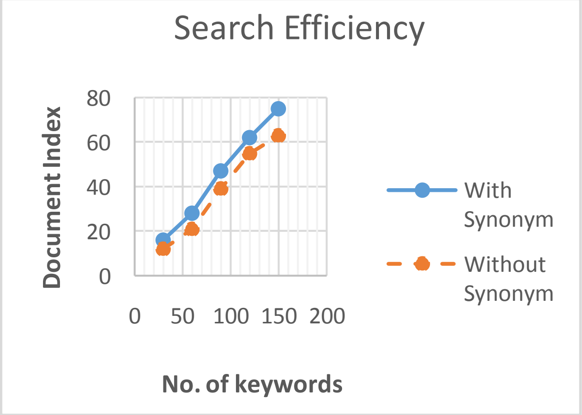


Fig.4 Search Efficiency

Search Efficiency is shown in the Fig. 4. It can be seen that number of indexes generated for the document will be more with synonym search when compare to keyword search without synonym. This increases the searching efficiency as we can search the documents with more coverage of its indexes.

V. CONCLUSION

In this paper, we considered the cipher text search in the scenario of cloud storage. We find out the problem of maintaining the semantic relationship between different plain documents over the related encrypted documents.

We propose the clustering architecture to adapt to the requirements of data explosion, and include the synonym search to increase the relevance between documents and relevance between the query and the documents.

In addition, we analyse the search efficiency and security by encrypting all the entities in the model such as indexes, keywords, synonyms and query keywords. Conducted experiments to evaluate the search efficiency and document relevance along with synonym search.

The experiment result proves that the proposed architecture not only properly solves the multi-keyword ranked search problem, but also brings an improvement in search efficiency, ranking of documents and the relevance between retrieved documents.

REFERENCES

[1] Chi Chen, Xiaojie Zhu, Student, Peisong Shen, Student, J.Hu, S.Guo, Senior, Z.Tari, Senior and Albert Y. Zomaya - An Efficient Privacy-Preserving Ranked Keyword Search Method - 2015
[2] Jin Li, Qian Wang, Cong Wang†, Ning Cao, Kui Ren and Wenjing Lou, "Fuzzy keyword search over encrypted data in cloud computing", Proceedings IEEE, 2010 - ieeexplore.ieee.org.
[3] N. Cao, C. Wang, M. Li, K. Ren, and W. J. Lou, ”Privacy-Preserving Multi-keyword Ranked Search over Encrypted Cloud Data,” in Proc. IEEE INFOCOM, Shanghai, China, 2011,pp. 829-837.
[4] W. Sun, B. Wang, N. Cao, M. Li, W. Lou, Y. T. Hou, and H. Li, ”Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking,” in Proc. ASIACCS, Hangzhou, China, 2013, pp. 71-82.
[5] Jarecki, S., Jutla, C., Krawczyk, H., Rosu, M., and Steiner, M. (2013, November). Outsourced symmetric private information retrieval. In Proceedings of the 2013 ACM SIGSAC conference on Computer and communications security (pp. 875-888). ACM.
[6] P. Golle, J. Staddon, and B. Waters, ”Secure conjunctive keyword search over encrypted data,” in Proc. ACNS, Yellow Mt, China, 2004, pp. 31-45.
[7] D. X. D. Song, D. Wagner, and A. Perrig, ”Practical techniques for searches on encrypted data,” in Proc. S & P, BERKELEY, CA, 2000, pp. 44-55.
[8] Deepa P L, S Vinoth Kumar, Dr S Karthik, searching techniques in encrypted cloud data, October 2012
[9] Y. C. Chang and M. Mitzenmacher, ”Privacy preserving keyword searches on remote encrypted data,” in Proc. ACNS, Columbia Univ, New York, NY, 2005, pp. 442-455. [10] Liu, C., Zhu, L., Li, L., Tan, Y.: "Fuzzy keyword search on encrypted cloud storage data with small index". In: Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference on, pp. 269–273. IEEE 2011
[11] Chuah, M., Hu, W.: "Privacy-aware bed tree based solution for fuzzy multi-keyword search over encrypted data". In: Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference on, pp. 273–281. IEEE 2011

DOI 10.17148/IJARCCE.2016.5858