AN EFFICIENT FRAMEWORK FOR SEARCH OVER ENCRYPTED DATA IN CLOUD

[1]Ms. Archana D. Narudkar, [2]Mrs. Aparna A. Junnarkar
Department of Computer Engineering,P.E.S Modern College of Engineering.Pune, India
Email:[1]archananarudkar@gmail.com, [2]aparna.junnarkar@gmail.com

Abstract
The cloud storage systems are most vulnerable to data security due to their internal data sharing among the servers. So data is always stored on the cloud by applying strong cryptographic techniques. Cloud is known for providing big storage capacity so performing search on huge encrypted data in cloud is posed as a real challenge. To solve the problem of searching many ideas were proposed which perform search over the encrypted data, but no system is providing complete accuracy as it mainly depends on the document content. In existing system they have used locality sensitive hashing as a search technique. In this paper we explores the existing technique of searching on encrypted data and also put forwards an idea of increasing speed of searching technique using Pearson correlation and Generalized inverted index.

Keywords : Correlation, AES,Generalised inverted index, feature data,trap door, Locality Sensitive hashing.

I. INTRODUCTION

The number of internet users across the globe increasing rapidly. Therefore use of cloud becomes a common practice nowadays. Due to complex computational structure of cloud and its data handling techniques it is unable to provide the security for all the stored data in the cloud.

As Cloud Computing becoming so famous and common, most of the sensitive information is being centralized on the cloud, such as emails, personal health records, government documents, etc. By storing their data into the cloud, the data owners can be relieved from the responsibility of data storage and maintenance so that they can enjoy the on-demand high quality data storage service. As owners of data and cloud server are not in the same trusted domain. It may put the outsourced data at risk, as the cloud server may no longer be fully trusted. Therefore it is a common practice to encrypt sensitive data usually prior to outsourcing for data privacy and combating unsolicited accesses. But encryption makes the search functionality a much more difficult task.

The common solution to search on encrypted data is after getting the user keyword for searching, every document needs to decrypt first and then the keywords need to match in every word of the document to retrieve the desired one[1]. But his process takes much more time to search the documents, so a need of proper and fast searching technique arises which can search the documents in the cloud without decrypting the data to save the cost of cloud service provider and time of the end users.

However, encryption of data makes data utilization a very challenging task given that there could be a large amount of outsourced data files. In Cloud Computing, there may be a sharing of outsourced data between data owners and a large number of users. The individual users might want to only retrieve certain specific data files they are interested in during a given session. One of the most popular ways is to selectively retrieve files through keyword-based search instead of retrieving all the encrypted files back which is completely impractical in cloud computing scenarios. Such keyword based search technique allows users to selectively retrieve files of interest and has been widely applied in plaintext search scenarios, such as Google search. Unfortunately, user's ability to perform keyword search on encrypted data is restricted and thus it is not possible to use the traditional plaintext search method for Cloud Computing.

To securely search over encrypted data, searchable encryption techniques have been developed in recent years. Searchable encryption schemes usually build up an index for each keyword of interest and associate the index with the files that contain the keyword. By integrating the trapdoors of keywords within the index information, effective keyword search can be realized while both file content and keyword privacy are well-preserved. Although allowing for performing searches securely and effectively, the existing searchable encryption techniques do not suit for cloud computing scenario since they support only exact keyword search. That is, there is no tolerance of minor typos and format inconsistencies. It is quite common that users' searching input might not exactly match those pre-set keywords due to the possible typos and user's lack of exact knowledge about the data.

In this paper we focused on the similarity search over the encrypted data technique which uses bloom filter and jaccard distance [2]. And also discusses the pros and cons of the similarity search method by analyzing its results by practically implementing the method.

The rest of the paper is organized as follows. Section II discusses related work and section III presents implementation and design of existing system and the design of our approach. Section IV gives the details of the results and some discussions we have conducted on this approach. Section V concludes the paper.

II. LITERATURE SURVEY

Cloud computing, a new terminology used to access data remotely from the centralized pool that can be rapidly deployed with great scalability and less computation overhead. Cloud computing comes with lots of advantages such as self-service on need, easy network access, accessing data independent of location, rapid resource elasticity, low pricing, transference of risk, etc. [3]. Thus we can say that cloud computing is advantageous to its user for avoiding large capital outlays required for deployment and management of software as well as hardware. Thus undoubtly cloud computing comes like a revolution in the field of information technology.

In this era of cloud computing, data owners get attracted to the cloud as it saves there lot of time, space and money. As the cloud computing starts gaining edge over another techniques of storing data, user starts storing large amount of information on cloud such as email ids, passwords, multimedia documents, companies secret data etc. by storing such information on cloud data owners get relief from the storing of their confidential data as cloud gives on demand access to the required things.

But the fact is that the data owners and cloud providers a may not be aware of each other regarding trust. So possibility of hacking or leaking confidential information is raises. This problem can be easily overcome by allowing the users to keep their information in encrypted format so that it is highly secure and chances of getting leaked is get minimized.

Cloud computing offers great data utilization, but searching over encrypted data is very challenging task as huge numbers of outsource files are presents there. However data owners might want selected files related with entered query. So keyword searching over encrypted data emerged as good technique to find the required data from the cloud.

Wei Zhou et al. in [4] proposed a new technique of searching which makes use of K gram technique for producing fuzzy logic results. Authors said that previous technique is good for searching when exact match of keyword is found. But if there are some spelling errors or some nearby words are entered then system will fails to give answers. Proposed system developed to overcome the said problem. It makes use of K gram based fuzzy logic to accomplish the task. For more security developers makes use of separate servers which are not related with each other's at all.

Jin Li, Qian Wang et.al. [5] Presents a real approach having fuzzy logic as base to search data over ciphered cloud data. Here author said that it is the very first approach for the same. Here system returns the file in which exact match of predefined word is get found. System also returns the file if it have close words from the entered query. To accomplish the task author edit distance is used, it helps in finding the similar words based on their semantics. At last authors comes on conclusion that it is the best system which uses fuzzy logic edit distance to solve the given problem. Wildcard-based technique which is a advanced methods is used by the authors to use fuzzy logic more efficiently

Prof. C. R. Barde [6] proposed a new technique known as Secured Multikeyword search (SMS) for searching the input query over ciphered data in cloud. To search files containing documents an efficient rule of coordinate matching is used. Coordinate matching refers to a technique of finding the huge number of matches as many as possible so that it will become easy to find similarity between the input query and the list of records. To improve further accuracy they developed an alert system, this system will gives an alert whenever an unauthorized users are trying to access the system. This alert is given by the emails and messages to the related users.

Ming Li et al. [7] Takes personal health records system (PHR) as a base for the case studies as it is the area which affected greatly due to the less security in cloud system. Here authors develop a framework known as Authorized Private Keyword Search (APKS) which is highly scalable for searching over cloud data. Further they make use of Hierarchical Predicate Encryption (HPE) to solve the problem of searching. Author declares that it is privacy search because it hides the keywords from the server which results in more and more security.

III. PROPOSED METHODOLOGY

In this section, we describe the method of implementation of similarity search technique on the encrypted data with the below mentioned steps as shown in Fig 1.

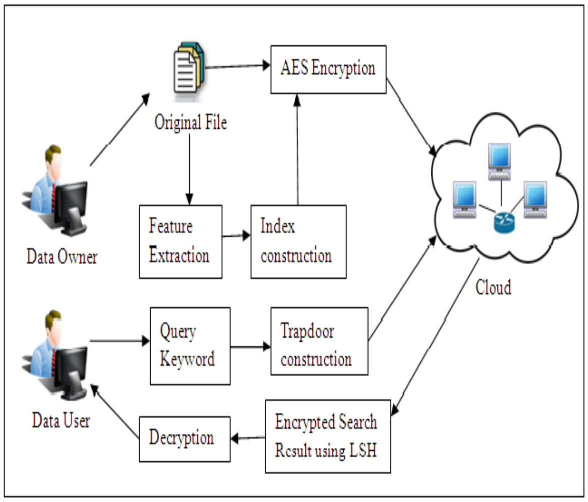

Fig 1: Overview of the Similarity search method

Step 1: From the original plaintext which data owner wants to upload the cloud, first features are extracted for the respective documents and preprocessing is done. Encryption of the data is done by using 256 bit AES encryption. And then data will be stored at cloud end.

Step 2: Searching process initiates from this step where user fires the query to get the desired searchable results for the uploaded data in cloud. On firing the user query the most important part of the similarity search method will trigger that is Locality Sensitive Hashing which can be describe with the below mentioned definition of the steps

Definition A - Bucket Construction: Here in this step matrix space translation is applied to create combination of words of the keywords which eventually enhances the process of similarity search. Then all this words are gathered in a vector container called as the Bucket.

Definition B – Trapdoor Creation: Here on each element of the bucket AES scheme is applied to get the collection of encrypted words which are the key matching substances with the stored encrypted data in the cloud.

Definition C – Bloom filter: To find the similarity searching some translation need to be done on the encrypted strings. And the simplest and effective translation can be done by embedding strings into bloom filters. Bloom filters are powered by the variable hashing schemes; in this case of similarity search bloom filters are bit array that are empowered with most powerful MD5 algorithm.

Definition D – Jaccard Distance: Once the strings are mapped into the bloom filters then the system uses the jaccard distance to measure the distances between the two set of strings as defined as follows.

$J_d (A , B)= 1 - | A \cap B | / | A \cup B |$

So at the end least distance sets are considered as the matched strings and finally string contained documents are returned to the user as the result of the searching process.

The Secure LSH index construction is summarized in Algorithm 1 as depicted in [2].

Algorithm 1: Build index

```
Require: D: data item collection,
 g: $\lambda$ composite hash functions,
$\Psi$ : security parameter,
MAX: maximum possible number of features

$K_{id} \leftarrow Keygen(\Psi)$, $K_{payload} \leftarrow Keygen(\Psi)$

for all $D_i$ € D do
    $F_i \leftarrow$ extract features of $D_i$
    for all $f_{ij}$ € $F_i$ do
        $f_{ij} \leftarrow$ apply metric space tranlation on $f_{ij}$
        for all $g_k$ € g do
            if $g_k(f_{ij})$ € bucket identifier list then
                add $gk(f_{ij})$ to the bucket identifier list
                initialize $V_{gk}(f_{ij})$ as a zero vector of size |D|
                increment recordCount
            end if
            $V_{gk}(f_{ij})[id(D_i)] \leftarrow 1$
        end for
    end for
end for
for all $B_k$ € bucket identifier list do
    $V_{Bk} \leftarrow$ retrieve payload of $B_k$
    $\pi_{Bk} \leftarrow Enc_{Kid}(B_k)$, $\sigma_{VBk} \leftarrow Enc_{Kpayload}(V_{Bk})$
    add $(\pi_{Bk}, \sigma_{VBk})$ to I
end for
return I
```

The similarity measure method never counts any scenario to low down the searching time process. So this may setback the performance of the similarity measure using locality sensitive hashing. So to increase the performance of the searching process we put forwards an idea of searching over the encrypted data using Pearson correlation method with generalized inverted index scheme [8] to fast the process of searching with much better accuracy. The approach of new idea can be seen in Fig 2.

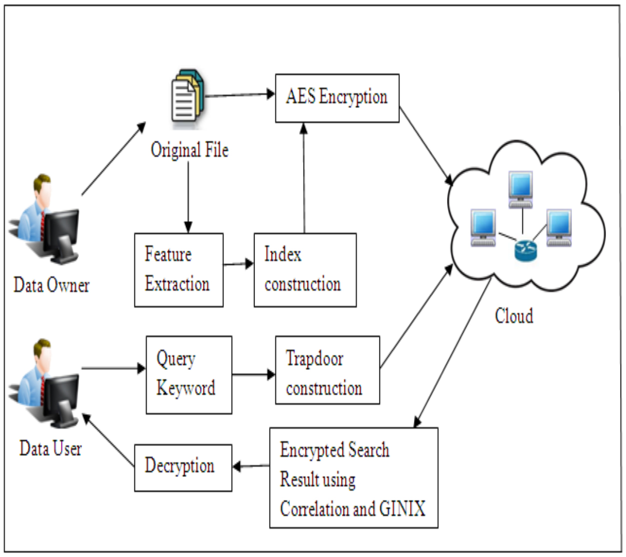

Fig 2: Overview of the proposed method

IV. RESULTS AND DISCUSSIONS

To show the effectiveness of proposed system some experiments are conducted on java based windows machine with Netbeans as IDE. To measure the performance of the system we set the bench mark for the number of retrieved documents.

To determine the performance of the system, we examined how many relevant documents are retrieved for the given keywords over the encrypted data in cloud.

To measure this precision and recall are considering as the best measuring techniques. So precision can be defined as the ratio of the number of relevant documents retrieved to the total number of irrelevant and relevant documents retrieved from the searching system. It is usually expressed as a percentage. This gives the information about the relative effectiveness of the system.

Whereas Recall is the ratio of the number of relevant documents retrieved to the total number of relevant documents not retrieved. It is usually expressed as a percentage. This gives the information about the absolute accuracy of the system.

Let we assign

- A = the number of relevant documents retrieved,
- B = the number of relevant documents retrieved are not retrieved, and
- C = the number of irrelevant documents retrieved.

So, Precision = (A/ (A+ C))*100

And Recall = (A/ (A+ B))*100

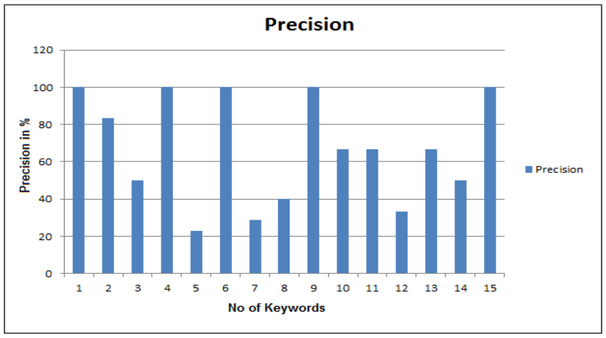

Fig.3. Average precision of the similarity search method

In Fig. 3, we observe that the tendency of average precision for the retrieved documents using similarity search method is about 68%.

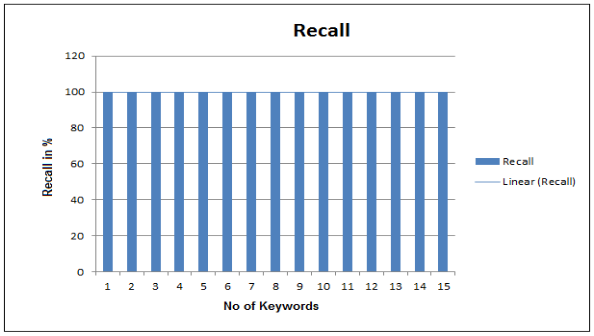

Fig.4. Average Recall of the similarity search method

In Fig. 4, we observe that the tendency of average Recall for the retrieved documents using similarity search method is 100%. This indicates similarity search gives more accuracy in terms of recall than the precision.

V. CONCLUSION

This paper thoroughly studies existing method of similarity search on the encrypted documents in cloud. And thereby analyze its results by implementing the same system which uses bloom filter and jaccard distance. The drawback of their system is that as it uses bloom filter it may give false positives. Also we found that much less work is done for the speed up of the process of searching. So we have given a system to lower the time for search on encrypted data. Our implemented system introduces with the Pearson correlation technique and generalized inverted index to reduce the searching time.

REFERENCES

[1] Ankit Doshi et.al. A Survey on Searching and Indexing on Encrypted Data. *International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 10, October - 2013*

[2] Mehmet Kuzu, Mohammad Saiful Islam, Murat Kantarcioglu. Efficient Similarity Search over Encrypted Data , *IEEE trasaction ,2012.*

[3] “Above the clouds : A Berkley view of cloud computing”

[4] Wei Zhou et al., K-Gram Based Fuzzy Keyword Search over Encrypted Cloud Computing. *Journal of Software Engineering and Applications, Scientific Research , Issue 6, Volume 29-32,January2013*

[5] Jin Li†, Qian Wang†, Cong Wang†, Ning Cao‡, Kui Ren†, and Wenjing Lou‡,”Fuzzy Keyword Search over Encrypted Data in Cloud Computing”*INFOCOM10.*

[6] Prof. C. R. Barde , Pooja Katkade , Deepali Shewale , Rohit Khatale . Secured Multiple-keyword Search over Encrypted Cloud Data.” *www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 2, February 2014)*

[7] Ming Li et al., Authorized private keyword search over encrypted data in cloud computing. *In: Distributed Computing Systems (ICDCS), 2011 31st International Conference on, pp. 383–392. IEEE (2011)*

[8] Hao Wu, Guoliang Li, and Lizhu Zhou. Ginix: Generalized Inverted Index for Keyword Search ,*IEEE Transcation, Volume 18, Number 1, February 2013 .*