International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 4, Issue 6 (Nov-Dec 2016), PP. 53-58

CO-RELATION TECHNIQUE FOR SEARCHING OF ENCRYPTED DATA OVER CLOUD

[1]Vrushali R. Charde , [2]Prof. Nitin S. More
[1,2] Dept. of Information Technology
Smt. Kashibai Navale College of Engg. Pune, India
[1]vrushali24692@gmail.com, [2] nsmore@sinhgad.edu

***Abstract*— Many data owners use cloud to outsource their data as cloud provides a platform to store a large amount of data which is very convenient and also reduces the data management cost. But, due to its complex computational structure and data handling techniques, the cloud is unable to provide the security for all the stored data. As well as the cloud storage systems are most vulnerable for the data security due to their internal data sharing among the servers. Therefore, by applying strong cryptography techniques, the data is stored in encrypted format over the cloud. But performing a search on this encrypted data is a real challenge as cloud provides a big storage capacity. Therefore, to solve this issue, many techniques are invented to perform the search on this encrypted data, but no method is providing complete accuracy because this mainly depends on the document content. The major drawback is all have a huge time complexity. Therefore, this paper approaches in improving the process of fast searching of encrypted data in the cloud. The proposed system puts forward an idea of correlation technique which is forwarded with inverted index and multi-parallel working threads.**

***Index Terms*— Cloud computing; trapdoor; searchable encryption; privacy-preserving; ranked search.**

I. INTRODUCTION

Cloud computing is an exponentially growing model as it provides a platform for the data owners to store their data from local sites to public cloud providing high flexibility and economic savings. Outsourced data ranges from health records, government documents, company finance data and e-mails. Also, the cloud can organize huge resource of computing, storage, and application, and enables users on-demand access to configurable computing resources with minimum economic overhead and with high efficiency. Therefore, both the individuals and the enterprises got motivated to outsource their personal and professional data over the cloud. Despite of these advantages, outsourcing sensitive data to the remote servers brings privacy concerns. To overcome this, sensitive data is usually stored in encrypted form to prevent from unauthorized access. So the cloud service providers manage to apply the cryptography algorithms to encrypt the data before storage process. And they provide original data to the users by using decryption technique on the same on their request. The users taking advantage of this, store an enormous number of documents which again creates the problem of searching the document in the cloud as all are present in the encrypted state.

The traditional solution to this is after getting the user keyword for searching; every document needs to get decrypted first and then the keywords need to match in every word of the document to retrieve the desired one. But this process takes much more time to search for the documents. So a need for proper and fast searching technique arises that can check the texts in the cloud without decryption the data to save the cost of cloud service provider and time of the end users. However, data encryption makes data utilization a very challenging task as there are a lot of outsourced data files. One of the most common techniques is to extract the files selectively through keyword-based search in spite of retrieving all the encrypted data. Such keyword-based search procedure allows the users to obtain the data of their interest selectively.

The rest of the paper is organized as follows: Section II summarizes the related work. Section III presents the system model. In Section IV results and discussion are described and Section V states the conclusion and future scope.

II. RELATED WORK

Cengiz Orencik and Erkay Savas [1], aimed to achieve an efficient system based on Private Information Retrieval (PIR) where any authorized user can perform their search on a remote database having multiple keywords that user is retrieving. The user can query the database provided that they possess trapdoor for the searched terms that allow the users to include them in their queries. This system is capable of performing searching for multiple keywords in a single query and gives the results so that the user can retrieve only the top matches. Cong Wang et al. [2] used build index along with Ranked Searchable Symmetric Encryption and the keyword frequency based relevance score. Here, order-preserving mapping scheme is used where small encrypted files are treated first, and then large encrypted files are processed. But system fails if multiple keywords are filled as input, with such input searching speed also increases. D. Boneh, G. D. Crescenzo et.al, [3] proposed a searchable encryption, where anyone having public key can write to the stored data in server but having restriction that only authorized users with private key can search. The disadvantage here is in the public key

setting, the privacy of keyword may possibly not be protected as the server has the ability to encrypt any keyword with the public key. As a result, it can be used to receive the trapdoor in order to evaluate the cipher text.

In order to retrieve a document containing only a word [4] describes a cryptographic model that solves problem of searching an encrypted data providing secure cryptosystem. It focuses on hidden query so that the untrusted server cannot search for a word without the user's authorization and it also support query isolation means, the server knows nothing more than the result giving an approach to search over remotely located data. Drawback here is it requires additional storage overhead and will not guarantees the security of the data. Eu-Jin Goh [5] defined a secure index with a security model for indexes known as a semantic security. An efficient Index Chosen Keyword Attack (IND-CKA) is constructed using pseudo-random functions and Bloom Filter. Where Hao Wu, Guoliang Li, and Lizhu Zhou [6], gives more efficient index structure called as Generalized INverted IndeX (GINIX) which combines consecutive IDs in an inverted list into a set of intervals to save the storage space. It also improved seeking performance of keyword as compared with the traditional inverted index. By re-ordering file in datasets, the GINIX method increases search speed.

Hongwei Li et al. [7] explained searchable encryption for multi-keyword ranked search on stored data. To develop an efficient multi-keyword search scheme k-nearest technique was used which returns the ranked searched results based on the accuracy. Where Jin Li et al. [8], proposed the main idea to form and solve the problem of fuzzy keyword searching over the encrypted cloud data while maintaining keyword privacy. Paper [9], Traffic and Energy saving Encrypted Search (TEES) architecture for mobile cloud storage applications has been introduced which achieves the efficiencies by employing and modifying the ranked keyword search as the encrypted search platform basis, reducing the energy consumption by 35~55 % by offloading computation of relevance scores to the cloud. Though TEES is implemented with security enhancement, but essential security defects of this approach is not completely resolved.Jin Li et al, [10] proposed fuzzy keyword search which focused on enabling effective as well as privacy preserving fuzzy keyword search in Cloud. It formalized the problem of effective fuzzy keyword search over encrypted cloud data by maintaining keyword privacy. Mehmet Kuzu et al. [11] introduced a method of locality sensitive hashing which is a high dimensional space searching technique .which uses a hashing technique to create trap door for searching encrypted documents in the cloud. Due to different cryptography methods, searchable encryption schemes can be constructed using public key cryptography or symmetric key cryptography.

Reza Curtmola et al. [12] used Searchable Symmetric Encryption (SSE) that allowed the third party to store its data on the server in a private manner. Also, multi-user SSE is constructed which is very efficient on the server side, on giving a trapdoor, the server only needs to evaluate a pseudo-random permutation to determine if the user is revoked or not. Here, only the owner of the data is capable of submitting search queries. [13]Address a content-based multimedia retrieval over encrypted databases which enables the client retrieval directly in the encrypted domain. Using scheme such as mini-Hash sketches and secure inverted index it uses jointly exploiting technique like cryptography, image processing, and information retrieval for securing indexes. This system is enhanced to overcome mini- Hash scheme that require longer sketches to obtain better performance. Y. Chang and M. Mitzenmacher [14] Makes use of PIR (Private Information Retrieval) queries for searching over the cloud. This method uses bloom filter gives storage space that can be useful to store some extra information. It hides the identity of the communication also keeps the semantic of the encrypted data. But it will not preserve the privacy and correctness of the data. Zhihua Xia et al. [15] introduced a ranked search scheme over encrypted cloud data using multi-keyword, where the greedy depth-first search algorithm is used to provide efficient multi-keyword ranked search. This scheme can flexibly achieve sub-linear search time and also deals with deletion and insertion of documents. The main drawback of this system is that it uses Symmetric Searchable Encryption scheme where the data owner himself needs to update information and send them to cloud server.

III. PROPOSED METHODOLOGY

This section describes the detailed implementation of a tool as follows.

A. System Framework

The overall working of the proposed method of Efficient Framework for Searching of Encrypted Data over Cloud can be described efficiently according to the steps which are described in below figure 1.

The input to the system is the original text file which is uploaded by the data owner and the output is the decrypted file which can be then accessed by the authorized user. In addition to the existing system [10], the proposed system uses Correlation technique and Generalized Inverted Index (GINIX) method for faster retrieval of data.

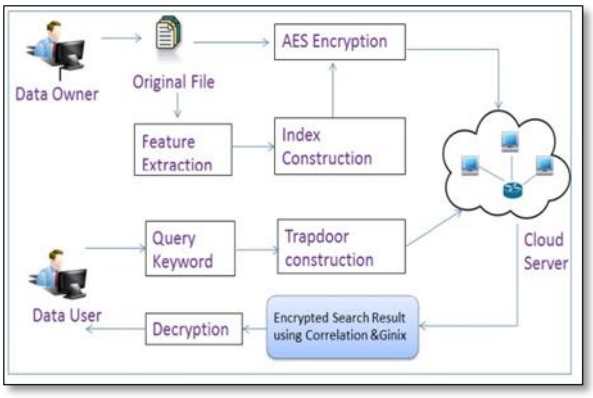

Figure 1: Proposed System Framework using co-relation search technique.

The two major functions of the system are:

1) While Uploading:

While uploading any plain text document over cloud, the following steps are performed on it:

a) Pre- processing :

In pre-processing, the input data is processed to produce output that is used as input to another program. Here, in the proposed system, pre-processing is done on the original input text so as to convert it into a basic format of words. Following two processes are applied on text:

a. Stop words: removes common words like 'a', 'an', 'the', 'who' etc.

b. Stemming: converts words into its basic form. Example- learning becomes learn

b) Index construction:

Index construction can also be called as extracting features. When the input data to an algorithm is too large to be processed, and it is suspected to be unnecessary, then it can be transformed into a reduced set of features (also named a "features vector"). This process is called feature extraction.

c) AES encryption:

Finally, on the extracted features, Advanced Encryption Standard Technique is applied. AES technique is used to convert plain text into encrypted text. And then this encrypted text is saved in another file over the cloud.

2) While Searching:

Once the file is uploaded in the encrypted format, the next step is to search for this encrypted file. Following are the steps performed by searching the encrypted file:

a) Query pre-processing:

While searching for any file over the cloud, the text for which the user is searching is also undergone from the pre-processing technique. Following two processes are applied to the text:

a. Stop words

b. Stemming

b) Trapdoor Creation:

Trapdoor creation is nothing but extraction of the featured which is converted into the encrypted format.

c) Bloom Filter:

After creation of trapdoor, to search for the trapdoor, Bloom filter is used. i.e. these words are in which file is examined here.

d) Co-relation :

This step is the major modification to the existing system [10]. A correlation illustrates a quantitative means of some type of correlation and dependence, indicating statistical relationships between two or more variables or observed data values. Pearson correlation coefficient, also known as r, R, or Pearson's r, a measure of strength and direction of the linear relationship between two variables that is defined as the covariance of the variables divided by product of their standard deviations.

e) Inverted Index:

Inverted index maps each word in the data-set, such as numbers or words, to its locations in a database file, or in the set of documents to a list of IDs of documents which appear in ascending order. The GINIX method [4] is used here where the main purpose is to provide fast full-text searches, at a cost of improved processing when a document is added to the database.

B. Algorithms

Algorithm 1 gives the steps to build the index for the documents while uploading the files. The input to this algorithm is the plain text which gives output as encrypted file with the indexes. All the features are extracted from the file and converted into secure index by using bucket identifiers.

Algorithm 1: Build index

Require: D: data item collection,
g: λ composite hash functions,
Ψ : security parameter,
MAX: maximum possible number of features
$K_{id} \leftarrow Keygen(\Psi)$, $K_{payload} \leftarrow Keygen(\Psi)$
for all Di ϵ D do
Fi ← extract features of Di
for all fij ϵ Fi do
fij ← apply metric space translation on fij
for all gk ϵ g do
if gk(fij) ϵ bucket identifier list then
add gk(fij) to the bucket identifier list
Initialize Vgk(fij) 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 ϵ 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 (π_{Bk} , σV_{Bk}) to I
end for
return I

Algorithm 2 gives the steps for searching the encrypted file over cloud. Here, by the index build on the file while uploading, the documents are searched over the cloud. While searching for any file, the keywords are compared with the contents of file using the Pearson co-relation technique and the most matching relevant files are given as output to user.

Algorithm 2: Searching algorithm
Step 0: Start
Step 1: Read the Query
Step 2: Create bucket B of query where minimum word of bucket should have three characters.
Step 3: Apply AES algorithm to encrypt the bucket.
Step 4: Read encrypted file from cloud
Step 5: divide content of file into words on space and store in a vector V
Step 6: for i=0 to length of B
Step 7: for j=0 to length of V
Step 8: if (V_j is equals to B_i)
Send length of B_i and Length of vector V to the Pearson correlation
Step 9: Set 0.5 as a threshold to get more precise files and store them in vector F
Step 10: end inner for
Step 11: End outer for.
Step 12 : Return File vector F

C. Calculations

This system uses following Pearson co-relation technique to find out the co-relation coefficient.

The formula is stated as below :

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sqrt{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} \sqrt{\sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}}}$$

r = Pearson Correlation Coefficient
n = Total No. of values in each Data set
X = the number of relevant documents retrieved
Y = the number of relevant documents retrieved and not retrieved, and
Z = The number of irrelevant documents are retrieved.
$\sum_{i=1}^{n} X_i Y_i$ = Sum of products of Paired Scores
$\sum_{i=1}^{n} X_i$ = Sum of X scores
$\sum_{i=1}^{n} Y_i$ = Sum of Y scores
$\sum_{i=1}^{n} X_i^2$ = Sum of Squared X squares
$\sum_{i=1}^{n} Y_i^2$ = Sum of Squared Y squares

It gives linear correlation between two variables X and Y, having value between +1 and -1. Depending upon the value of correlation coefficient r, the matching relevant file correlated to keywords searched can be determined.

For high co-relation, value of r ranges from 0.5 to 1.0, for medium co-relation ranges from 0.3 to 0.5 and for low co-relation it ranges from 0.1 to 0.3

IV. RESULTS AND DISCUSSIONS

Some experimental evaluations are performed to show the effectiveness of the system. And these experiments are conducted on the windows based java machine with universally used IDE Net-beans. Also, the numbers of retrieved documents are used to set the benchmark for performance evaluation. Numbers of relevant documents retrieved from the private cloud for the set of keywords are used to show the effectiveness of the system.

Numbers of scenarios present where one measuring parameter dominates the other. By taking such parameters into consideration, two measuring parameters such as precision and recall are used.

Below are the definition of the used measuring techniques i.e. precision and recall.

Precision: it is a ratio of numbers of proper documents retrieved to the sum of total numbers of relevant and irrelevant documents retrieved. Relative effectiveness of the system is well formulated by using precision parameters.

Recall: it is a ratio of total numbers of relevant documents retrieved to the total numbers of relevant documents not retrieved. Absolute accuracy of the system is well described by using recall parameter

Therefore,
Precision = (X / (X+ Z))*100
And
Recall = (X / (X+ Y))*100

TABLE I: Precesion and Recall Chart

T=Number of document extracted by the system	X= Relevant Documents in T	Y=No. of relevant documents not extracted	Z= Number of irrelevant Documents extracted	Precision = (X/ (X+ Z))		Recall = (X/ (X+ Y))	
				Correlated Search	Fuzzy Search	Correlated Search	Fuzzy Search
2	1	0	1	0.5	0.66	1	0.93
3	2	0	1	0.666666667	0.65	1	0.94
6	5	0	1	0.833333333	0.67	1	0.92
4	4	0	0	1	0.7	1	0.93
6	6	0	0	1	0.75	1	0.95
2	2	0	0	1	0.75	1	0.9

For the above values of X,Y and Z, on calculating the precision and recall values using the formula, we get the following graphs :

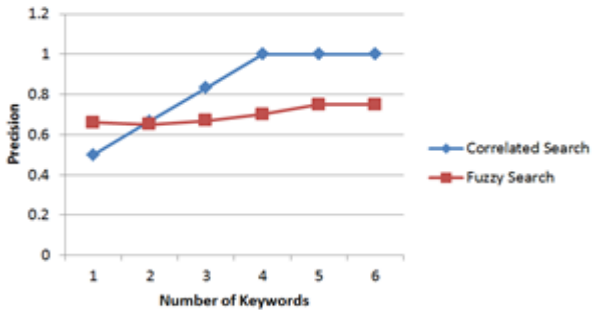

Figure 2: Average precision comparison by co-related search and fuzzy search method.

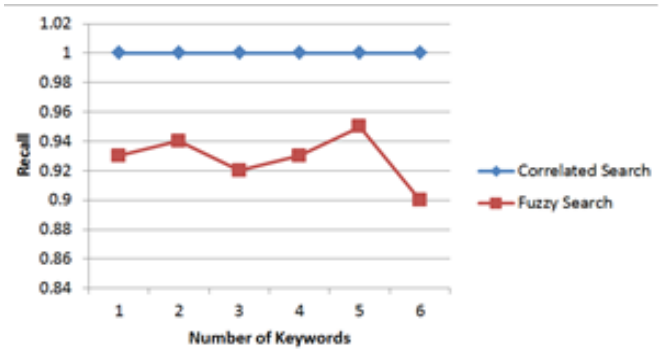

Figure 3 : Average Recall by co-related search and fuzzy search method.

In Figure 2, it is clear that the average precision obtained by using similarity search method is approximately 88%. Figure 3 shows that the system gives 100% recall for the co-relation search method. By comparing these two graphs, we can conclude that the co-relation search method provides high recall value compare to the precision value.

We observe that the tendency of average Recall for the retrieved documents is about 1 which is better than the fuzzy Search Method.

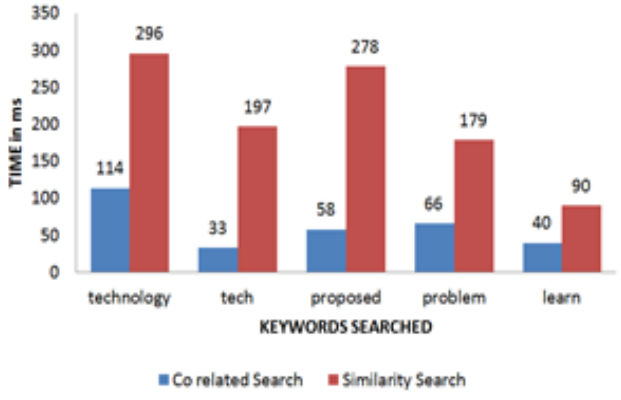

Figure 4 : Results obtained when searched for the keywords like technology, tech, proposed, problem, learn using co-relation search and similarity search technique, depending on time parameter.

IV. CONCLUSION AND FUTURE SCOPE

The proposed system uses co-relation technique and inverted index model to handle huge data for searching in cloud based on the extracted features of the original data. This system allows searching for data in the cloud that is that is in encrypted format in such a way that data should not be decrypt for searching process. Due to this process, the time of searching can be saved efficiently.

As all the systems give importance to security and time parameters of the system, future work can be reducing the energy consumption of the system. Also, System can be enhanced to implement in Internet of things paradigm and can enhance to work in all format of data.

REFERENCES

[1] CengizOrencik and ErkaySavas, " Efficient and Secure Ranked Multi-keyword Search on Encrypted Cloud Data ", ACM 978-1-4503, March 2012

[2] Cong Wang et al.," Enabling Secure and Efficient Ranked Keyword Search over Outsourced Cloud Data", IEEE Transactions on Parallel and distributed systems, vol. 23, no. 8, August 2012

[3] D. Boneh, G. D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search," in Proc. of EUROCRYPT, IEEE Conference on Computer Communications 2004.

[4] D. X. Song, D. Wagner and A. Perrig," Practical Techniques for Searches on Encrypted Data"., Proceedings of the 2000 IEEE Symposium on Security and Privacy, pp. 44-55.

[5] E.-J. Goh et al ,"Secure indexes." IACR Cryptology e-Print Archive, vol. 2003, p. 216, 2004.

[6] Hao Wu.Guoliang Li, and Lizhu Zhou, "GINIX : Generalized Inverted Index For Keyword Search", Tsinghua sciences and technology,ISSN,Vol.18.

[7] Hongwei Li, DongxiaoLui, Yuanshun Dai, "Enabling Efficient Multi-keyword Ranked Search Over Encrypted Mobile Cloud Data Through Blind Storage", IEEE Transactions, vol. 3, no. 1, Mar. 2015.

[8] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, "Fuzzy keyword search over encrypted data in cloud computing", in INFO COM, 2010 Proceedings IEEE. IEEE, 2010, pp. 1–5.

[9] Jian Li, Ruhui Ma, and Haibing Guan, "TEES : An Efficient Search Scheme over Encrypted Data on Mobile Cloud ", IEEE transactions, vol. 3, no. X., 2015

[10] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, "Fuzzy keyword search over encrypted data in cloud computing," in Proc. of IEEE INFOCOM'10 Mini-Conference, San Diego, CA, USA, March 2010.

[11] M. Kuzu, M. S. Islam, and M. Kantarcioglu, "Efficient similarity search over encrypted data," in Data Engineering (ICDE), 2012IEEE 28th International Conference on. IEEE,2012,pp.1156–1167.

[12] R. Curtmola, J. Garay, S. Kamara, and R. Ostrovsky, "Searchable symmetric encryption: improved definitions and efficient constructions," in Proceedings of the 13th ACM conference on Computer and communications security. ACM, 2006, pp. 79–88.

[13] Wenjun Lu, Ashwin Swaminathan, Avinash L. Varna, and Min Wu, "Enabling search over encrypted multimedia databases," in proc. of SPIE Media Forensicsand Security,09,2009.

[14] Y.-C. Chang and M. Mitzenmacher , "Privacy preserving keyword searches on remote encrypted data," in Proceedings of the Third International Conference on Applied Cryptography and Network Security. Springer-Verlag, 2005, pp. 442–455.

[15] Zhihua Xia, 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, Vol.no. 2015.