International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Mining of Encrypted Data over Cloud Data

Aashish Arjun Bugalia†*, Abhishek Kumar Mishra†, Preeti Naresh Ajwani†, Karan Shah Arora† and Shalini V. Wankhede†

†Department of Computer Engineering, Sinhgad Academy of Engineering, SPPU, Pune, Maharashtra-411048, India

Accepted 28 Nov 2015, Available online 03 Dec 2015, **Vol.5, No.6 (Dec 2015)**

Abstract

Due to the increasing popularity of cloud computing, more and more data owners are motivated to outsource their data to cloud servers for greater convenience and reduced cost in data management. However, sensitive data should be encrypted before outsourcing for privacy requirements, which obsoletes data utilization like keyword based document retrieval. For searching the encrypted data has to be decrypted first and then searched, which consumes too much time and resources. So we need an idea of searching over the encrypted data without decrypting it thus enhancing the process of searching. This is achieved by extracting data features while uploading the file into cloud and then conducting search on these features of data to get the relevant files. We are using a bucket data structure to store these data features. The AES encryption scheme is used to encrypt the index and query. We ensure accurate relevance score calculation between encrypted index and query vectors through Pearson's correlation.

Keywords: *Cloud Computing, Web Server, Client/ server, Distributed Database, Distributed Networks, Encryption.*

1. Introduction

The cloud storage systems are most vulnerable for security due to their internal data sharing among the servers. To overcome this, data is always stored in the cloud by applying strong cryptographic techniques. But eventually this doesn't solve the problem of storing process as cloud is known for providing big storage capacity, so performing search on this huge encrypted data in cloud poses a real challenge.

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.

2. Literature Review

(Song *et al*, 2012) first gave the practical solution for the searchable encryption on data. They used word by word document encryption. In this server is given the capability to search on encrypted document. They focused more on keyword based search. The disadvantage of this scheme is it is not secure, computation complexity is linear to the whole collection length and memory overhead is large.

*Corresponding author: **Aashish Arjun Bugalia**

(Swaminathan A *et al*, 2012) first time attempted secure rank-ordered search. For that for each document in set they collected term frequency information for building indices. To protect against statistical attacks they secured the indices. Depending on encrypted queries it ranks the documents and document having most rank will be pushed up using ranked method. The cryptographic techniques such as order-preserving encryption is applied by using term frequencies and other document information. It calculates the relevance score for every document and identifies documents which are most relevant. The given method is well suited for large documents and also it provides higher accuracy and security. But for this method computational cost is high and protecting communication link is bit difficult task and combating traffic analysis is not done.

(Wenjun Lu *et al*,2010) Addresses, content-based multimedia retrieval over encrypted databases that enable client retrieval directly in the encrypted domain. Firstly feature vectors are extracted and vocabulary tree is used to cluster them hierarchically. Indexing is done based on vocabulary tree which is represented as a bag of visual words. It describes how many times the representative feature vectors in the vocabulary tree occur in the questioned image. In order to secure index scheme such as mini-Hash sketches and secure inverted index, it uses jointly exploiting technique like cryptography, image processing, and information retrieval. First schema exploits randomized hash functions and the second

schema makes use of inverted indexes of visual words. This model is further en- handed to overcome mini-Hash scheme that require longer sketches to achieve better performance in order to achieve performance similar to that of the inverted index scheme. The approach used is different slightly in terms of feature. Query does not consider specific feature for an image but all features of an image.

(Qin Liuy et. al, 2009) proposed secure and privacy preserving keyword search(SPKS). Their solution is practical and efficient. In this scheme without leaking any information about plaintext the computational overhead on users is reduced by participating cloud service providers in partial decipherment. This scheme is provably secure and without being aware of any keyword and email information it enables cloud service provider (CSP) for the determination of whether the keyword is present in given email.

(Ming Li *et al*, 2000) considered authorized private keyword search problem on encrypted data in cloud computing. They took the scenario where encryption of records and keyword index is done by multiple data owners so that multiple users can search. They proposed an authorization framework where search abilities to users are obtained from local trusted authorities according to their attributes.

(Mehmet Kuzu,2007)in proposed an approach which uses Locality Sensitive Hashing (LSH) which is a nearest neighbor algorithm for the index creation In this similar features are hashed to the same bucket with a high probability due to the property of LSH while the features which are not similar reside in the different buckets. If the dataset is small then the communication cost and search time required by this scheme is better. But if the database size increases the time required for communication and for search also increases rapidly. In this method every query feature increases linearly with the size of database. They used bloom filter for translation of strings. It is a data structure which gives information about whether an element is present in the collection or not. But the disadvantage of this structure is that it is a probabilistic data structure. The false positive rate is more in this structure. It shows if keyword definitely not in the set of encrypted documents or may be present in the encrypted collection of documents. Also Jaccard distance is used by them to measure the distance between strings.

(Jianfeng Wang *et al*, 1996) Presents an important approach that not only guarantees the confidentiality and security but also the verifiability of the searching method. They used Verify () algorithm which outputs true if pass else false. They used edit distance for measuring the similarity between two strings.

3. Problem Statement

In the current scenario, the time required to search a file on the cloud is high. So we need to create an efficient search retrieval process on cloud files that will reduce time in a considerable way. The basic idea of searching system over encrypted data in cloud comes from the fact that the data in cloud is having poor security norms. So data needs to be always stored in encrypted format while storing. To search the required data by the user on the encrypted data ,requires data to be decrypted first and then search, so this eventually slows down the process of searching. So to overcome that problem we search over the encrypted data without decrypting the original data which enhances the process of searching and reduces time complexity.

4. The Definitions

4.1. Pearson's Correlation Coefficient

The use of the Pearson Correlation Coefficient as a means of reducing noise during speech processing is vast . While the Mean Square Error Criterion (MSE) is a fairly unique method to identify noise, the Signal to Noise Ratio behaviour which is an important aspect of noise reduction cannot be identified by the MSE, thereby making the use of the Pearson Correlation Coefficient highly imperative. Analysis of noise reduction performance is comparatively, more favourably accomplished by using the Pearson Correlation coefficient. Pearson Correlation Coefficient (PCC) utilizes the normality of variables that are analysed . The quality or how accurately the variables can be related to each other defines PCC. It makes use of quantitative variables. Pearson's Correlation Coefficient assumes that the variables will always have a linear relationship which have to be measured on normal scales. The relation between Pearsons Correlation Coefficient and Saltons cosine measure is based on numerous divisions of the norm of a vector. A threshold value can be given for the cosine set by devising an algorithm above which none of the corresponding Pearson correlations would be of a negative value.

4.2. Bloom Filter

Bloom filter has been used as an important method to solve many internet problems . Though they are effective, they are prone to vulnerabilities. Encoding is done via bloom filter which involves source routing of the protocols. It elaborates on the use of Counting Bloom Filters and introduces a method to improve the efficiency of Counting Bloom Filters(CBF). Though they consume significant amount of memory, they are used as widely as Bloom filters due to their fast set representations resulting in a lesser amount of errors, thereby supporting membership queries and element deletions.

4.3. AES Encryption

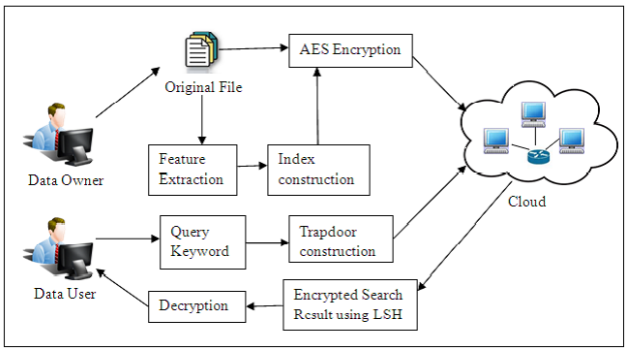
AES Encryption has been used as an effective method to encrypt data. It presents various advantages compared to other encryption techniques such as Reverse Cycle Cipher whose execution time derails the

mechanism. Modification of the AES to improve its performance is done with the help of neural network tool that could prove to be fairly advantageous. The AES is a 128 bit based Encryption technique and the design of this system has been programmed by Very High Speed Integrated Circuits. Various implementations of AES are possible when testing its use on a multi core system. Sixteen such implementations of the AES are mapped by exploring different levels of parallelism. The system contrives a design that produces a considerably larger amount of throughput per unit of chip area. It brings about changes in efficiency of the AES when compared with the original standard.

5. Methodology

5.1. Proposed System

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.



Step 1: From the original plaintext which data owner wants to upload on the cloud, firstly 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 on cloud. On ring the user query the most important part of the similarity search method which will trigger that is Locality Sensitive Hashing which can be described 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 enhance the process of similarity search. Then all these 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

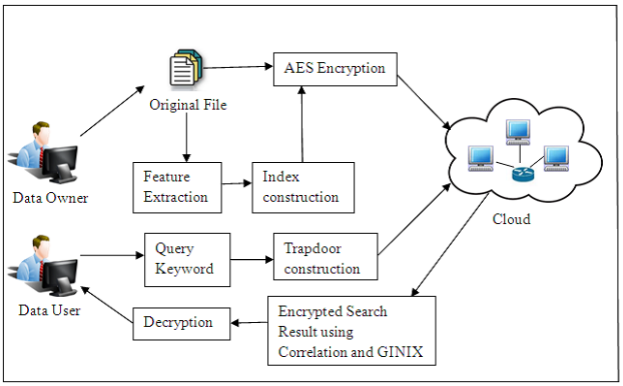
To find the similarity searching some translation needs to be done on the encrypted strings. And the simplest and effective translation can be done by embedding strings into bloom filters. Bloom filters are powered by the variable hashing schemes; in this case of similarity search bloom filters are bit 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.

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

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 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 forward an idea of searching over the encrypted data using Pearson correlation method with generalized inverted index scheme to fasten the process of searching with much better accuracy. The approach of new idea can be seen in Fig 2.



5.2. Advantages

1. Increased security using Encryption and can work efficiently on BigData.
2. Decreased time of searching due to no decryption required during searching process.
3.Effective keyword search used with both file content and keyword,privacy are well preserved.

5.3. Limitations

1. Advanced security issues are not properly handled.

2. Accountability of the operations over data is not very efficient.
3. Difficult future work to design a dynamic searchable encryption scheme whose updating can be completed by cloud server only.
4. Dishonest data user may distribute his/her secure keys to the unauthorized ones which may cause security issues.
5. Due to no decryption sometimes it is difficult to extract data.

Conclusion

Our interest in Cloud Computing was a driving force behind the selection of this topic. As mentioned, we have studied different techniques and we have come to the conclusion that AES Encryption and Bloom Filter surpass other techniques in having very few drawbacks but a high quality output is ensured by these algorithms. The file search on Cloud will be done in a considerably lesser amount of time that will be noticed by end users. By eliminating decryption during searching of any data on cloud, the searching time can be massively reduced while security breaches can be prevented. Newer and better algorithms such as Bloom Filter can be made use of for cost effectiveness.

References

M. Kuzu, M. Islam, S.Mohammad, and M.Kantarcioglu (2012) Efficient similarity search over encrypted data. In *Proceedings of the IEEE 28th International Conference on Data Engineering, pages 1156-1167.*

Qiang Tang(2012) Search in Encrypted data: Theoretical Models and Practical Applications

C. Wang, N. Cao, J. Li, K. Ren, and W. Lou (2010), Secure ranked keyword search over encrypted cloud data, in *Proc. of ICDCS10*, pp.253 262.

J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou(2009), Enabling efficient fuzzy keyword search over encrypted data in cloud computing, *in Cryptology ePrint Archive, Report 2009/593.*.

D. Song, D. Wagner, and A. Perrig (2000), Practical techniques for searches on encrypted data, in *Proc. of the IEEE Symposium on Security and Privacy 00, pp. 4455*

H. Park, B. Kim, D. H. Lee, Y. Chung, and J. Zhan (2007), Secure similarity search, *in Cryptology ePrint Archive, Report 2007/312.*

O.Goldreich and R. Ostrovsky,(1996) Software protection and simulation on oblivious rams, *Journal of the ACM, vol. 43, pp. 431473.*