ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Efficient Semantic Search over Encrypted Data in Cloud Computing

Yash Khare[1], Asif Shaikh[2], Akshay Gole[3], Prof. K. Bala[4]

B.E. Student, Department of Computer Engineering, BVCOEL, Pune, India[1]

B.E. Student, Department of Computer Engineering, BVCOEL, Pune, India[2]

B.E. Student, Department of Computer Engineering, BVCOEL, Pune, India[3]

Professor, Department of Computer Engineering, BVCOEL, Pune, India[4]

ABSTRACT: Now a day, use of cloud storage for storing and retrieving huge amount of data has been increased tremendously. Storing and retrieving such a large amount of data consumes lot of time as data in the cloud needs to be always stored in encrypted format while storing and needs to be decrypted while searching. This ultimately slows down the process of searching. To avoid this massive consumption of time, data searching speed can be increased by directly searching over encrypted data in the cloud. There are many methodologies available in the market for searching the encrypted data over the cloud. Most of these methodologies for search over encrypted data in the cloud are having some performance issues that can lead to lower the accuracy of the technique, so this paper proposes a novel idea of search over encrypted data in the cloud. This method applies the data search over encrypted data in the cloud and this is achieved by extracting data features while uploading the file into a cloud and search is conducted on this features data to get the appropriate files. In this paper, we are implementing bucket creation technique along with AES encryption for storing data to the cloud. In addition to this we are using a bloom filter for searching necessary information over the encrypted data in the cloud. The main advantage of our methodology over the existing search systems is that it supports an extreme speed of data searching.

KEYWORDS: AES encryption, bloom filters, cloud, feature extraction, index construction, trapdoor construction.

I. INTRODUCTION

This basic idea of searching system over encrypted data in the cloud comes from the fact that the data in the cloud are having poor security norms. So, data need to be always stored in encrypted format while storing. To search the required data by the user for the encrypted data requires data to be decrypted first and then search, which eventually slow-down the process of searching. In order to overcome this an idea of searching required data over the encrypted data without decrypting the original data enhances the process of searching.
Given fig. (A) is the data flow diagram of search over encrypted data.

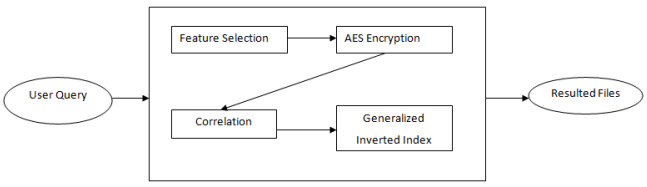

Fig. (A)

There are four methods which play an important role in search over encrypted data. They are:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

A. Feature Selection

Feature selection, also known as attribute selection, is the processof selecting a subset of relevant feature for use in model construction. Feature selection is used for shorter training time and for simplification of model for easy interpretation.This technique ids often used in domains where there are many features and few data points.

B. AES Encryption

AES or Advanced Encryption Standard is a method for encrypting and decrypting information or data. Whenever you transmit files over secure file transfer protocols like FTPS, SFTP, WebDAV's, OFTP or HTTPS, there's a good chance your data will be encrypted by AES 128, 192, or 256.

The algorithm of AES operates on plaintext block of 16 bytes. Encryption of shorter block is possible by padding the source bytes with null bytes. This can be accomplished by several methods, the simplest of which assumes that the final byte of the cipher identifies the number of null bytes of padding added. This algorithm is not used only for secured service, but for fast, reliable and it is easily implemented in both software and hardware.

C. Correlation

Correlation describes the relationship between attributes and datasets. Correlation defines similarity among the attributes. A correlative standard specifies how a set of nodes in a facility graph, when selected for the same service composition, may impact the QoS weights on the outgoing edges from these nodes. A service graph may have a set of correlative criteria.

D. Inverted index

An inverted index is also known as postings file or inverted file. It is an index data structure for storing a map from content, such as words or numbers, to its locations in a database file, or in a document or a set of documents. The inverted index is used to allow the full text search when a document is added to the database. There are two main variants of inverted index for storing the data. They are recorded level inverted index and word level inverted index. The record level inverted index contains a reference to a document for each word and word level inverted index contains the position of each word.

In this paper section 2 represents related work and section 4 elaborate proposed techniques in detail. The performance of the systems analysed in section 4 and finally this paper is concluded with the future extension traces in section 5.

II. RELATED WORK

1. Feature Selection describes the process of attribute selection. For achieving accuracy in goals feature selection selects a smallest necessary set of features [1]. In data mining, Feature selection is one of the simple phases of pre-processing [2]. The main aim of feature selection is to improve the detection speed and accuracy as well as reduce dimensionality of data.

2. AES Encryption introduces the symmetric encryptionalgorithm. AES is based on principle of substitution-permutation network. It provides a method for encrypting and decrypting data. AES is highly secured and fastest encryption and decryption technique [3]. Client side encryption is an operative approach to deliver security to transmitting data and stored data [3]. AES performs all its computations on bytes. The AES algorithm has 3 fixed 128-bit block ciphers with cryptographic keys i.e. 128 bits, 192 bits and 256 bits, the size of the key is unlimited, where the block size is maximum 256 bits, AES encryption technique is fast, flexible and secured [3].

3. A correlative standard specifies how a set of nodes in a facility graph, when selected for the same service composition, may impact the QoS weights on the outgoing edges from these nodes. A service graph may have a set of correlative criteria [4]. Correlation describes the relationship between attributes and datasets. Correlation is determined similarity among the attributes [5].

4. Generalized inverted index defines the list of information. Inverted lists are typically used to index fundamental documents to retrieve documents according to a set of keywords efficiently [6]. Inverted index is one of the greatest

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

efficient searchable index structures and has being generally adopted in plain text search [7]. Inverted index is the best popular data structures used in document retrieval systems [7].

III. PROPOSED METHODOLOGY

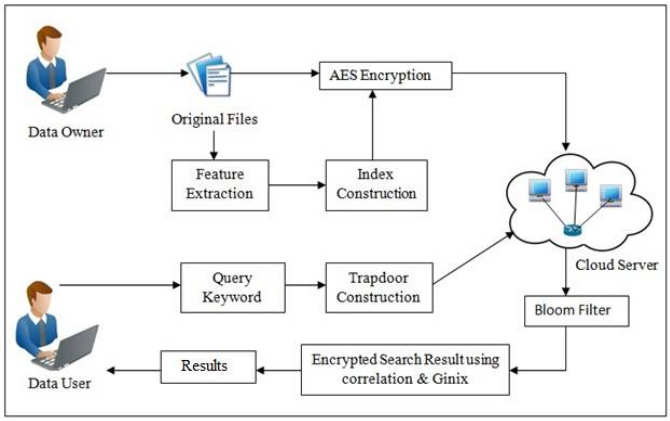

Fig. (B)

Given fig. (B) shows the working model of searching over encrypted data.

Step 1: Uploading of Data
Initially Data in text file format is being uploaded by data owner. Subsequently three sub process occurs. Initially original data is being encrypted using AES and pattern buckets are Stored in Encrypted format in Cloud. Secondly Data is being pre-processed,matrix Translation is being done creating buckets,data is being stored in as feature data in private cloud. An original copy of data is being stored as complete data in other cloud.

Step 2: Input Query pre-processing
Input Query is being submitted by Data user to System. System Accepts query performing pre-processing on query. Future query is being Trapdoor to create Query words pattern in encrypted format. Every word is being sent matrix translation creating buckets, this bucket is future encrypted using AES Algorithm.
Her in This process pattern are being generated using bucket generation. A bucket is set of word list for every word starting from trigram to N gram Approach.
Bucket generation Process is as below

“Computing Generates bucket as shown below

“Computing”➔{ com,comp,compu,comput,computi,computi,computin}

Step 3: Bloom Filter Application
Data in cloud Stored as feature data is being dynamically copied in feature vector consisting of File name and File Feature. This complete Data is termed as Bloom Filter Data.

Step 4: Search correlation
Search correlation Pattern matching is being done with Input Query and Trapdoor query . Both vectors are being compared and Correlation vector for input query is being generated. This vector is being sent for future correlation evaluation.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Step 5: Ginix index
Search results are being displayed in all files that have being matched for given query input

---

Algorithm 1: Build index

Require: D: data item collection,
Ψ : security parameter,
MAX: maximum possible number of features
$K_{id} \leftarrow Keygen(\Psi)$, $K_{payload} \leftarrow Keygen(\Psi)$
for all Di € D do
Fi ← extract features of Di
for all fij € Fi do
fij ← apply metric space translation on fij
for all gk € g do
if gk(fij) € 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 € 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: MATRIX SPACE TRANSLATION

//Input: Data collection Set D ={D_i}
//Output: Matrix space Set MS
Step 0:Start
Step 1: Get the Set D
Step 2: FOR i=0 to Size of D
Step 3: get $S_i$of Di
Step 4: FOR j=0 to length of si
Step 5: sb_i=substring(s_i,2→j)
Step 6: Add sb_i to MS
Step 7: END FOR
Step 8: END FOR
Step 9: return MS
Step 10: Stop

IV. RESULTS

Some trial assessments are performed to demonstrate the adequacy of the framework. What's more, these analyses are led on windows based java machine with generally utilized IDE Net beans. Likewise, the quantities of recovered records are utilized to set benchmark for execution assessment. Quantities of applicable recovered archives from the cloud for the arrangement of catchphrases are utilized to demonstrate the adequacy of the framework. The following are the meaning of the utilized measuring methods i.e. exactness and review.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Exactness: it is a proportion of quantities of appropriate archives recovered to the total of aggregate quantities of important and immaterial records recovered. Relative viability of the framework is very much communicated by utilizing accuracy parameters.

Review: it is a proportion of aggregate quantities of applicable archives recovered to the aggregate quantities of pertinent records not recovered. Supreme precision of the framework is all around described by utilizing review parameter

Quantities of situations presents where one measuring parameter rules the other. by thinking about such parameters, we utilized two measuring parameters, for example, accuracy and review.

For Detailed Examination
• A = No. of relevant docs retrieved
• B =No of relevant documents not retrieved
and
• C = No of irrelevant documents are retrieved.

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

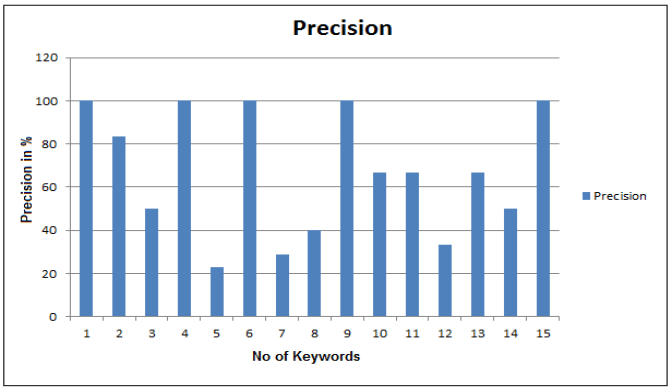

Fig.3. Average precision of the similarity search method

In Fig. 3, by observing figure 3 it is clear that average precision obtained by using similarity search method is approximately 65%.

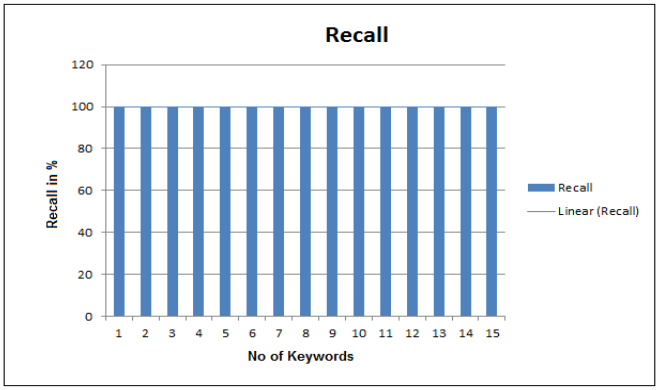

Fig.4. Average Recall of the similarity search method

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

In Fig. 4, figure shows that the system gives 95% recall for the similarity search method. By comparing these two graphs we can conclude that the similarity search method gives high recall value compare to the precision value.

V. CONCLUSION

Search over encrypted System assist in retrieving File that contains required information without decryption process. This would Enhance search process in cloud where files are b being encrypted and stored. In above project work search time has being found to be low. effective retrieval of documents measuring precision and recall.
Future system could be evaluated using Pearson correlation for similarity search.

REFERENCES

[1] Yogesh Dhote, Shikha Agrawal, and Anjana Jayant Deen, "*A Survey of Feature Selection Techniques for Internet Traffic Classification*", International Conference on Computational Intelligence and Communication Networks, Volume 7, December 2015.
[2] Mehdi Barati, Azizol Abdullah, Ramlan Mahmod, and Norwati Mustapha, "*Features Selection for Ids in Encrypted Traffic Using Genetic Algorithm*", International Conference on Computing and Informatics, ICOCI, Volume 04, pp 038, August 2013
[3] Mr. Niteen Surv, Mr. Balu Wanve, Mr. Rahul Kamble, Mr. Sachin Patil, and Mrs. Jayshree Katti, "*Framework for Client Side AES Encryption Technique in Cloud Computing*", IEEE International Advance Computing Conference (IACC), 2015.
[4] Jun Huang, Shihao Li, Qiang Duan, Ruozhou Yu, and Shui Yu, "*QoS Correlation-Aware Service Composition for Unified Network-Cloud Service Provisioning*", IEEE, December 2016.
[5] N P Nethravathi, Prasanth G Rao, and Indramma M BMSCE, "*CBTS: Correlation Based Transformation Strategy for Privacy Preserving Data Mining*", IEEE international WTE Conference on Electrical and Computer Engineering, Volume 157, Number 1, December 2015.
[6] Hao Wu, Guoliang Li, and Lizhu Zhou, "*Ginix: Generalized Inverted Index for Keyword Search*", Tsinghua Scienceandte Chnology, ISSNll1007-0214ll10/12llpp77-87, Volume 18, Number 1, February 2013.
[7] Bing Wang, Wei Song, Wenjing Lou, Y. Thomas Hou, "*Inverted Index Based Multi-Keyword Public-key Searchable Encryption with Strong Privacy Guarantee*", IEEE Conference on Computer Communications, May 2015.
[8] Kaitai Liang, Xinyi Huang, Fuchun Guo, and Joseph K. Liu, "*Privacy-Preserving and Regular Language Search over Encrypted Cloud Data*", IEEE Transaction on Information Francis and Security, Volume 11, Issue 10, October 2016.

BIOGRAPHY

Yash Khare is pursuing his B.E Degree from Bharati Vidyapeeth's College of Engineering, Lavale, Pune, India. He is being affiliated to Savitribai Phule Pune University, Pune, India. He is currently pursuing his B.E Degree in Computer Science and Engineering. His current research interest includes Artificial Intelligence, Network Security, and Algorithm Design.

Asif Shaikh is pursuing his B.E Degree from Bharati Vidyapeeth's College of Engineering, Lavale, Pune, India. He is being affiliated to Savitribai Phule Pune University, Pune, India. He is currently pursuing his B.E Degree in Computer Science and Engineering. His current research interest includes Operating Systems, Network Security and Networking.

Akshay Gole is pursuing his B.E Degree from Bharati Vidyapeeth's College of Engineering, Lavale, Pune, India. He is being affiliated to Savitribai Phule Pune University, Pune, India. He is currently pursuing his B.E Degree in Computer Science and Engineering. His current research interest includes Network Security and Networking.

Prof. K. Bala has completed her B.Tech. Degree in Computer Engineering from Ghanashyam Hemalata Institute of Technology and Management, Puri, Odisa, India and pursued her degree from Biju Patnaik University of Technology, Odisha, India. She later on completed her M. Tech Degree in the field of Computer Science from C.V. Raman College of Engineering, Bhubaneshwar, India and pursued her degree from Biju Patnaik University of Technology, Odisha, India. She is currently working as a professor for Computer Engineering Department in Bharati Vidyapeeth's College of Engineering, Lavale, Pune, India.