International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 2, February 2016

An Efficient Searching Method over Semantically Secure Encrypted Data

[1]Swathy J, [2] Surya S R
[1] PG Scholar, [2] Assistant Professor
[1] Department of Computer Science, [2] Department of Information Technology,
College of Engineering, Perumon, Kerala, India
[1]lakshmiuma91@gmail.com, [2] sooryasr07@gmail.com

Abstract- **Data mining has wide variety of real time application in many fields such as financial, telecommunication, biological, and among government agencies. Classification is the one of the main task in data mining. For the past few years, due to the increment in various privacy problem, many conceptual and feasible solution to the classification problem have been proposed under different certainty prototype. With the increment of cloud computing user have an opportunity to offload the documents and processing to the cloud, in an encrypted form. The documents in the cloud are in encrypted form, existing privacy preserving classification systems are not relevant. The privacy preserving classification to preserve security of documents, protects the user query, and hides the access mode. Using PPKNN classification efficiently searching encrypted document in the cloud.**

Index Terms— Encryption, k-NN Classifier, Outsourced Databases, Security

I. INTRODUCTION

Data mining is a powerful new technique to discover knowledge within the large amount of the data. Also data mining is the process of discovering meaningful new relationship, patterns and trends by passing large amounts of data stored in corpus, using pattern recognition technologies as well as statistical and mathematical techniques. These data are may be numbers, or sequence of characters that can be processed by a computer.

Now a day's cloud computing model [1] is changing the structure of the organizations' way of storing, accessing, and processing their data. As the growing processing data, many organizations to focus on the cloud computing in terms of its efficiency, flexibility, security, and document control .To reduce the data overhead the companies offload their data to the cloud. Most often, organization give their computational activity in addition to their data to the cloud. For all massive advantages of cloud computing provide, privacy and security problems in the cloud are blocking organization to use this advantages. The data are in encoded form, traditional encryption scheme execute any data mining function over encrypted data is very challenging process. There are other privacy problems shows by the following example.

Example: Assume a business organization offloads its encrypted important document set and appropriate data mining task to the cloud. Business organization to find one document for their project. The organization uses the classification method to find that document. First the organization wants to create document record x (query) contains the document relevance. Then this query x can be send to the cloud, and the cloud will calculate the class label for x and retrieve the relevant document. Since x hold sensitive details to protect the document privacy and x should be encrypted before outsource to the cloud.

The above example reveals that the Data Mining over Encrypted Data (denoted by DMED) cloud also want to protect the users' document when the document is the part of a data mining process. Also cloud can prove useful and sensitive documents about the original documents by detect data access patterns even if the documents are in encrypted format. For the privacy and security reasons DMED issue have main three reasons (1) security of the encrypted documents (2) security of the users query record and, (3) conceal document access pattern.

The classification is the main task in data mining. The privacy preserving data mining operations (Classification/Clustering etc.) has thus become an important problem in current years. Each classification has its own advantages, is provide privacy preserving k-Nearest Neighbor classification over encrypted document in the cloud computing environment is a challenging problem. This efficient Privacy Preserving k-Nearest Neighbor (PPkNN) classification technique is used for searching over encrypted documents. The classification and searching are

performed only over that encrypted documents therefore not reveal any information about those documents.

II. LITERATURE SURVEY

In the past, traditional encryption schemes were used for data security. Here the client sends both the cipher text and the private key to the server. The server then decrypts and processes them by using the client's private key. If the client's private key is compromised, any intruder can easily access the sensitive data [2]. This encryption is not applicable in cloud environments, as the cloud is an open structure, so that any number of intruders can gain access to the private data. As a result, the traditional methods fail in providing a better security. In 1979, Shamir introduced the first secret sharing scheme [3] consisting of following steps,

1. Choose any prime number pmax(s; n+1). Let Zp represents the field of integers modulo p.
2. Choose a1, a2,……ak-1ϵ Zp, randomly, uniformly, and independently.
3. Let q(x) = D+a1x1+a2x2+……. +ak-1xk-1
4. Let Di=q (i), 1≤i≤n. (The evaluation of q (i) done over Zp)

This scheme was not secure against intruders. So in 1998, Thomas J introduced an efficient and accurate secret sharing method [4]. The method worked on the following basis: Assume that probability of undetected cheating is less than ϵ, for any ϵ > 0.

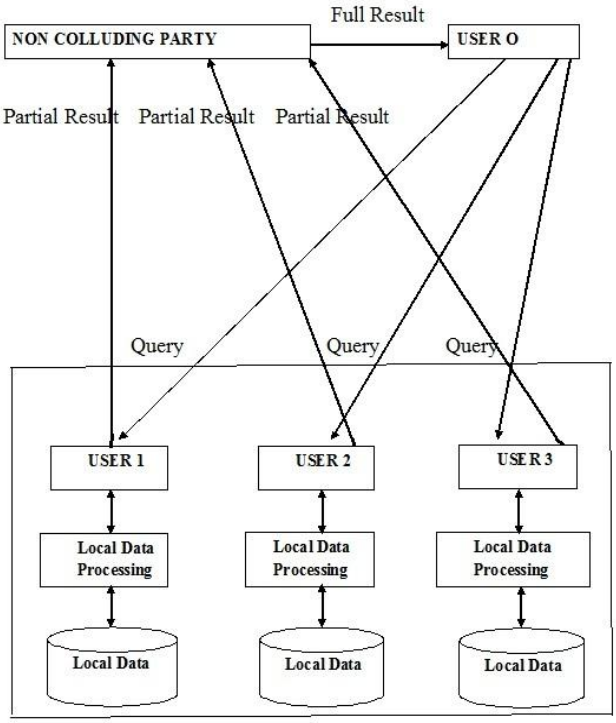
1. Choose any prime number pmax ((s - 1), (k- 1)/ ϵ =+ k, n).
2. Choose a1, a2,……..ak-1 ϵ Zp, randomly, uniformly, independently.
3. Let q(x) = D+a1x1+a2x2+……..+ak-1xk-1
4. Choose (x1, x2……..., xn) uniformly and randomly from all permutation of all distinct elements from 1, 2, 3……p-1. Let Di = (xi, di), where di=q (i).

Suppose the k - 1 participant is intruders, these methods cannot handle the existing security problems. The marten van Dijk and Ari Juels are prove traditional encryption schemes are not solve the privacy preserving problem in cloud computing environment [5]. In 2009, Craig Gentry introduced a Fully Homomorphic encryption scheme [6] aiming for a better level of security. The scheme could evaluate the circuit over encrypted data without employing any decryption process. The system works using the following steps:

(1) Using encryption to evaluate the arbitrary circuit.
(2) Using same encryption to find its own decryption circuit.
(3) Using encryption scheme to find its decryption circuit boots trappable

Even though the method is more secure than normal encryption schemes, it is very expensive and their usage in real time applications has not yet been explored fully. Homomorphic encryption serves as a cheap alternative which makes it an apt choice to be used in various real world applications. Homomorphic encryption is to provide security for large data. Current methods for privacy preserving data mining [7] cannot be applied to Data Mining over Encrypted Data problems (DMED). This problem can be solved using a Secure Multi-party computation model [8]. In this method, sensitive data is collected and processed and also they provide an efficient general purpose computation system to address this issue. Share mind is a virtual machine for privacy-preserving data processing that depends on the share computing method.

Nowadays, there exist different methods of classification techniques [9] in data mining. Each classification technique has its own merits and demerits. In 2006, Murat Ksantarcioglu and Chris Clifton introduced privately computing k-nearest neighbor classification in a distributed manner [10]. If any user want to find the class label for an attribute x, the user sends a query to the each of the systems so that each system individually calculates the k-nn based on Euclidean distance between attributes and x. After processing each user sends the partial result to non colluding third party. The outsourced partial result is encrypted using the user's public key. Therefore the third party cannot access the result. Third party to calculate the final class label based on the partial result. User accepts the final result from the non-colluding third party and decrypts it. This method use the normal encryption scheme for security, there for the data are not that much secure. The architecture of this system is shown in Fig. 1



1: Information flow in secure k-nn classifications

Provide privacy preserving k-nearest neighbour classification [11] is to use the semantically secure homomorphic encryption scheme. The privacy preserving k-nearest neighbour classification technique can classify large amounts of data in a much secured fashion. In this technique, user gives a query xq wants to compute the distance between xq and each training instances. Each user handle only portion of the instances they compute the distance measure according to their attribute values. Find the k-nearest neighbour of xq all users to sum their distances together. To obtain the resulting distance without losing data privacy, one user to generate multiple queries and collect the distance (Euclidian) from each query and to find the final result.

III. PROBLEM DEFINITION

Suppose Alice owns a file D it contains n sensitive documents. Alice wants to store this sensitive document to the cloud because of the limited storage capacity of the systems. Initially, Alice to encrypt the documents using fully homomorphic encryption [12] this encryption technique is semantically secure. Let the encrypted file denoted by D' and Alice to outsources D' to the cloud for further classification and searching process. Initially cloud to classify documents based on the Privacy Preserving k Nearest Neighbor classification then to apply the searching over that classified documents.

Let Alice want to search a sensitive document in the cloud. She sends an input query x (testing data) to the cloud and gets the relevant documents.

$$PPkNN (x, D') \rightarrow Document$$

IV. PROPOSED SYSTEM

In this section propose an efficient searching method over semantically secure encrypted documents. As mentioned earlier assume that Alice have document set D it contains n sensitive documents. Alice wants to store these sensitive documents to the cloud because of the limited storage capacity of the systems. Initially, Alice to encrypt the documents using fully Homomorphic encryption this encryption technique is semantically secure. Let the encrypted file denoted by D' outsourced to the cloud for further classification and searching process. The architecture of the system is shown the Fig 2

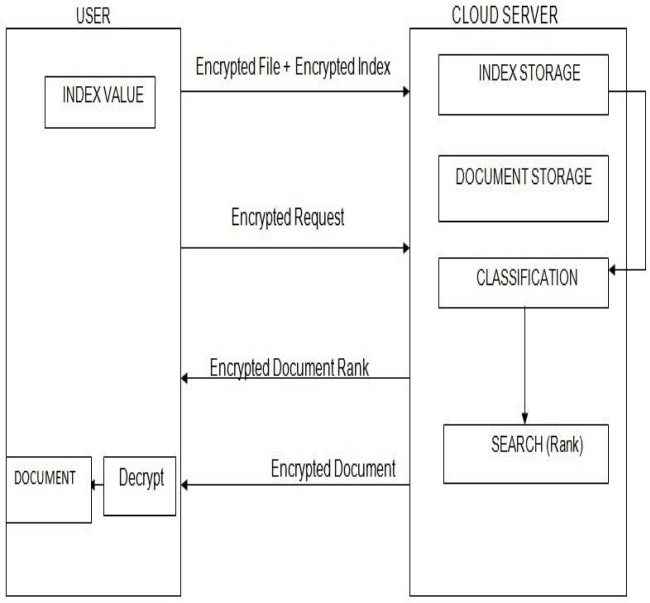


Fig 2: System Architecture

In this protocol the user sends both encrypted file and index to the cloud server. Documents are encrypted using fully homomorphic encryption. User first to extract the N words from the documents based on the word frequency threshold. Here to ignore words less than word frequency threshold. User encrypts each word frequency (Document relevance), file name and index value then send to the cloud in the matrix form as shown in the Fig 3

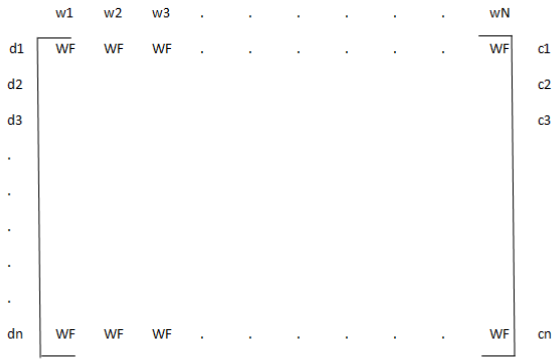


Fig 3: Encrypted Input Matrix (D')

In the above matrix {d1, d2……dn} is represents the n documents and {w1, w2,……..wN} is the filtered meaningful words. Each documents have a class label (class value) is represents the set {c1, c2……cn}. WF is the word frequency is also in encrypted form therefore any outside user cannot understand the relevance of words in any documents. Cloud

server accepts the input matrix and store document and index to corresponding storages.

Server classify the documents (testing documents) based the following Privacy preserving k Nearest Neighbor (PPkNN) algorithm.

Algorithm 1: PPkNN (x, D') → Document
Initialize Dist [L], L- Document size, List

1. for each document di in D
   Dist [i] = Euclidian Distance (Epk(x), Epk (di))
2. for each document di
   If (size (List) < Nk) Add di to List Else Any document in the List have distance greater than Dist [i], if so remove that document and add di to List.
3. Initialize Class [Nc] = 0
4. For each document in the List
   Class [di .Class] ++
5. Find Class having maximum value selects that class

User give an encrypted query x to the server. Server fined the Euclidian distance between the query x and documents then store these distances to matrix Dist [L]. In step3 and step4 for finding the relevant documents by create a documents list add minimum distance documents to that list using distance matrix. From that list select the maximum value class and get the relevant documents.

V. CONCLUSION

Different type of privacy preserving classification has been introduced in past few years. This method is not applicable to outsourced documents. An efficient privacy preserving k-Nearest Neighbor classification over encrypted document in the cloud and is used to preserve security of the documents, security of user query, and hide the access pattern. Efficiently perform a searching over encrypted document using the privacy preserving k-nearest neighbor (PPkNN) classification. This protocol provides complete security to bulk of documents.

REFERENCES

[1] .F, "Data Mining: Concepts, Models, and Techniques," Springer, 2011.

[2] P. Mell and T. Grance, "The nist definition of cloud computing (draft), "NIST special publication " vol. 800, p. 145, 2011.

[3] Subodh Gangan, "A Review of Man-in-the-Middle Attacks".

[4] A. Shamire, "How to share a secret" Commun. ACM, vol. 22, pp. 612-613, Nov. 1979.

[5] IBM Thomas J, "How to share secret with cheaters" Journal of Cryptology, I: 133-138, 1988.

[6] Martenvan Dijk and Ari Juels, "On the Impossibility of Cryptography alone for Privacy-Preserving Cloud computing".

[7] C. Gentry, "Fully homomorphic encryption using ideal lattices" in ACM STOC, pp. 169-178, 2009.

[8] R.Natarajan1, Dr.R.Sugumar, M.Mahendran and K.Anbazhagan, " A survey on Privacy Preserving Data Mining " International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1, Issue 1, March 2012.

[9] Y. Lindell and B. Pinkas, " Privacy preserving data mining " in Advances in Cryptology(CRYPTO), pp. 36 - 54, Springer, 2000.

[10] Thair Nu Phyu, "Survey of Classification Techniques in Data Mining " Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol IIMECS 2009, March 18 - 20, 2009.

[11] M. Kantarcioglu and C. Clifton, "Privately computing a distributed k-nn classifier in PKDD", pp. 279 - 290, 2004.

[12] Jiadi Yu, Peng Lu, Yanmin Zhu,Guangtao Xue, and Minglu Li Toward Secure Multikeyword Top-k Retrieval over Encrypted Cloud Data, IEEE transactions on dependable and secure computing, vol. 10, no. 4, july/august 2013.