Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 8, August 2013, pg.224 – 229

RESEARCH ARTICLE

Privacy Control Methods for Anonymous & Confidential Database Using Advance Encryption Standard

Madhuri Patil[1], Sandip Ingale[2]
[1]Dr. Babasaheb Ambedkar University, Aurangabad, India
[2]Dr. Babasaheb Ambedkar University, Aurangabad, India

[1] *patil.madhuri167@gmail.com*, [2] *sadip.support@gmail.com*

Abstract— Many Government agencies , business and organizations are willing to collect large amount of data containing the sensitive information about individual such as income, diseases & symptoms also wants to release or share that data to other parties for research work. Thus release of such data with sensitive information (microdata) violates individual's privacy. To protect sensitivity or confidentiality of shared data, it often needs to be sanitized before it can be distributed and analyzed. A popular and effective method for sanitizing data is called data anonymization. Data anonymization is the process of replacing the contents of identifiable fields (such as IP addresses, usernames, Social Security numbers and zip codes) in a database so records cannot be associated with a specific individual, project or company. There are various anonymization techniques that can be used such as Data encryption, randomization, etc. In this paper we observe various privacy conserving techniques with their advantages as well as disadvantages also presents suppression & Generalization based approach for privacy preserving simultaneously propose a system "Privacy conserving of anonymous & confidential Database using AES approach" that are more strong to achieve privacy.

Key Terms: - Privacy preservation; anonymization; confidentiality; suppression; Generalization

I. Introduction

Now a day's database serves as a valuable asset for many applications in various fields like medical research, Intelligence agencies, Bank etc. Thus today security of database is becoming very important issue. There are huge numbers of database containing numerous sensitive information (microdata) such as PIN, account number, income, diseases etc., if all this information falls into wrong hands then it would be very dangerous. So there is a big concern of privacy.

Privacy is the right of individual to keep their information secret hiding from others [5].privacy & confidentiality often used as a synonyms but there is vast difference between them. Privacy relates to person & confidentiality relates to nondisclosure of certain information to anyone except authorized person. For example in case privacy Health history or exam results discuss in private area, may include asking an accompanying family members or friends to leave the room temporarily , whereas in case of confidentiality patient information is not shared with other research team members. Therefore confidentiality is also a big issue. Here we provide a survey on various privacy preserving approaches with their merits & demerits [1].There are lots of techniques developed for privacy conservation, but one well known technique is K- anonymization.

Anonymization technique enables transferring between two organizations by converting text into human readable form using encryption method [1].Anonymization is about preserving identifying or private information using encryption. Nonanonymized version of data deleted from sender side after it is being sent to the receiver side. This is one of the important concepts in this technique [2].Such technique provide security by modifying data in such a way that it gives the same results for more than two tuples. When provider wants to insert a tuple into database, then it gives birth to two problems concerning both the individual's privacy as well as confidentiality of database. i) Is updated database still privacy preserving. ii) Does database owner needs to know the data to be inserted.

II. SURVEY OF PRIVACY PRESERVING METHODS

A huge number of techniques have been developed to provide privacy to the database such as, cryptographic approach, bucketization, Randomization, & K-anonymity.

A) cryptographic approach

The cryptographic approach for privacy conserving data mining assume that The cryptographic approach for the privacy preserving data mining is assume that the data is stored at several private parties and they accept the describe the result of specific data mining operation. The parties use a cryptographic protocol for encrypting and decrypting the messages. That is they use encrypted messages to do some operation efficient. They blindly run their algorithm These mining process could be occurred in between two untrusted parties, or even between competitors The main target is to protecting privacy in distributed mining process and to perform this data mining process two different approaches are available these are partitioned the data horizontally and that on vertically this method gives perfect, secure result. But it is very slow efficient.

Advantages

1. It is very effective method for protecting the privacy
2. It uses a different cryptographic algorithm for improving efficiency.

Disadvantage

1. This method becomes complicated when more than few parties involved.

B) Bucketization

Bucketization removes the identifiers from the data and also partitions tuples into buckets. Buckets contain the subset of tuples. Generalization transforms the QI values in each bucket into "less specific but semantically consistent" values. So that tuples of the same bucket cannot be distinguished by their QI values. In bucketization separates the SAs from the QIs but randomly permuting the SA values in each Bucket.

Disadvantages

1. It does not prevent membership disclosure.
2. It requires a clear separation of QI attributes and the sensitive attribute.

C) Randomization

Randomization is an effective way which prevents the user from learning sensitive data which can be easily implemented because the noise added to the given record is independent from the other records. The amount of noise is large enough to smear original values, so individual record cannot be recovered. The randomization method is simple as compare to other methods because it does not require to knowledge of other records. Large randomization increases the uncertainty and user's personal privacy. They claim that approaches may lose information as well as not provide privacy by introducing random noise to the data by using random matrix properties, [13]. It successfully separates the data from the random noise and subsequently discloses the original data.

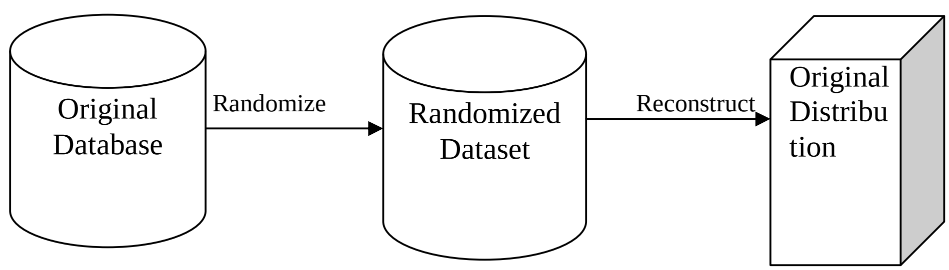

Figure 1 Randomization Method

Advantages

1. The randomization method is very simple method and which can be easily applied when we collect the data.
2. It is support to protect individuals' privacy.
3. Due to its simple working anyone can use and it is more efficient.

Disadvantages

1. It is not suitable when multiple attribute databases are used.
2. It is very slow technique because when data collector collects the data from data provider the data provider adds some noise in data and to reorder that data it takes more time.

D) Anonymity

Anonymization means identifying information is removed from the original data to protect personal or private information. There are many ways to perform data anonymization basically this method uses k-anonymization approach if each row in the table cannot be distinguished from at least other k-1 rows by only looking a set of attributes, then this table is K-anonymized on these attributes [7].

Example: If you try to identify a person from a table, but the only information you have is his birth date and gender. There are k people meet the requirement [1][3][4].

To understand the concept of data anonymization considers a simple example of medical patient. The information of a single patient is stored in a single line i.e. tuple, and database is store confidentially at the server. The users may be medical researchers who have the access to DB. Since DB is anonymous, one main concern is to protect the privacy of patients. Such task is guaranteed through the use of anonymization. if the database DB is anonymous, it is not possible to identify the patients record.

There are two mostly used anonymization techniques as follows

1) Suppression based K-anonymization

The basic concept of suppression based algorithm is to mask some attributes by special value *,In this algorithm t stands for tuple which is to be inserted by data provider & T stands for the Anonymous Database. QI stands for Quasi-Identifier which consists of set of attributes that can be used with certain external information to identify a specific individual.

The Suppression algorithm works as follows.

Step 1: User X sends User Y an encrypted version containing only the s non-suppressed attributes.

Step 2: User Y encrypts the information received from User X and sends it to her, along with encrypted version of each value in his tuple t.

Step 3-4: User X examines if the non-suppressed QI attributes is equal to those of t. If true, t can be inserted to table T. Otherwise, when inserted to T, t breaks k- anonymity.

2) Generalization Based K-anonymization

In generalization algorithm are replaced with more general values based on value hierarchy Garph (VGH) [9].

The protocol works as follows:

Step 1: User X randomly chooses a δ ε *Tw* (Witness Set).

Step 2: User X computes γ = GetSpec (δ).

Step 3: User X and User Y collaboratively compute s= SSI (γ, τ).

Step 4: If s=u then t‟s generalized form can be safely inserted to T.

Step 5: Otherwise, User X repeats the above procedures until either s=u or witness set is empty.

Table 1: Original Dataset

Disease	Gender	Age
Typhoid	Girl	19
HIV	Man	50
Typhoid	Woman	45
Exothrix	Man	68
HIV	Boy	20
Exothrix	Woman	63

Table 2: Suppressed Data with K=2

Disease	Gender	Age
*	Girl	*
*	Man	*
*	Woman	*
*	Man	*
*	Boy	*
*	Woman	*

Table 3: Generalized Data with K=2

Disease	Gender	Age
Disease caused by bacteria	Female	[1-30]
Disease caused by Virus	Male	[31-60]
Disease caused by bacteria	Female	[31-60]
Disease caused by Fungi	Male	[61-100]
Disease caused by Virus	Male	[1-30]
Disease caused by Fungi	Female	[61-100]

Table 1 represents the actual information in the original Dataset, after applying the suppression based algorithm over the original dataset the original dataset is anonymized & displays anonymized records. It makes changes in two QI & hence value of K=2 in Table 2.Table 3 shows the results of Generalized method with replacing the value after the data mining process is applied. The “Data Mining” point can be generalized to more specific value with “Database System”.
So, by applying this concept & replacing the remaining values in table with more general value the original dataset is anonymized using generalized method & finally when T is K-anonymous, we can delete duplicate tuples, & we call the resulting set the witness set of T[1].

Merits:

i) By replacing actual value with more general value it become very difficult to find or guess actual data.
ii) K-anonymous techniques is very fact and efficient as compared to previous techniques.
iii) By replacing actual value with “*”unauthorized user get confused and it creates more possible combination related to original dataset.

Demerits:

i) The main problems with generalization are it fails on high-dimensional data due to the curse of dimensionality it causes too much information loss due to the uniform distribution assumption.
ii) The database with the tuple data does not be maintained confidentially [5]

III. Proposed System of k-Anonymization using AES Technique

Anonymization is about preserving identifying or private information using encryption. The main purpose is to protect sensitive information. In a k-anonymous dataset, if any identifying information is found in the original dataset with k tuples then first we identifies quasi-identifiers i.e. the tuple that clearly distinguish the given tuple in database. Then we apply suppression based algorithm, in this algorithm we are identifying quasi-identifier & we are computing a K-partition which is a collection of disjoint subset of rows in which each subset contains at least K rows & the union of these subset id the entire table, then we are replacing each records with ‘*****’.In suppression based algorithm we are using diffie Hellman Key exchange algorithm [10] to generate private secure key. Then we are applying the AES (Advance Encryption Standards) algorithm [10] to encrypt & decrypt data by using the key generated by diffie Hellman key exchange algorithm. In this approach we are dealing with encrypted data not with original data. When user enters their information, then we encrypt it by using AES algorithm simultaneously we also encrypt the data in table using same algorithm. If information inserted by user matches with the table, then tuple will be decrypted & inserted into table. Generalization based Approach we are replacing the value in table with the more general values. If the data entered by the user matches with the value being replaced by the general value then this record will replaced by the general value and these general values being inserted into table.

Overview of Proposed System in the figure 2 compares existing data updates and make sure there is no redundancy and helps to analyze the data in database. K-Anonymization allows database to maintain a suppressed and generalized form of data such that data is much secured. The cryptography technique is used to secure the saved data in database safely such that the information is encrypted, stored and can be retrieved and decrypted back to original with specific authorization.

The Diffie Hellman Key exchange algorithm:

Public key encryption scheme based on a commutative encryption function.

1. Alice encrypts message M with her key: $ka \rightarrow \{M\}ka$.
2. Alice sends {M}ka to Bob.
3. Bob in his turn encrypts the received message: $\rightarrow \{\{M\}ka\}kb$
4. Bob sends {{M} ka} kb back to Alice.
5. Alice is able to decrypt the received message due to commutatively

$\{\{M\}ka\}\text{kb} = \{\{M\}kb\}ka :\rightarrow \{M\}kb$

6. Alice sends {M}kb to Bob, who can decrypt it using his key $kb \rightarrow M$

Diffie and Hellman use a commutative encryption function based on discrete logarithm:

Appropriate prime p and generator g are chosen, and common for all users.

1. Alice chooses a secret random number xa(→her private key) and publish $ya = g^{xa}$ (her public key).
2. Bob does the same with xb secret and $yb = g^{xb}$ public.
3. Alice uses $yb^{xa} = g^{xa\,xb}$ to encrypt a message to Bob.

Bob uses $ya^{xb} = g^{xa\,xb}$ to decrypt the received message.

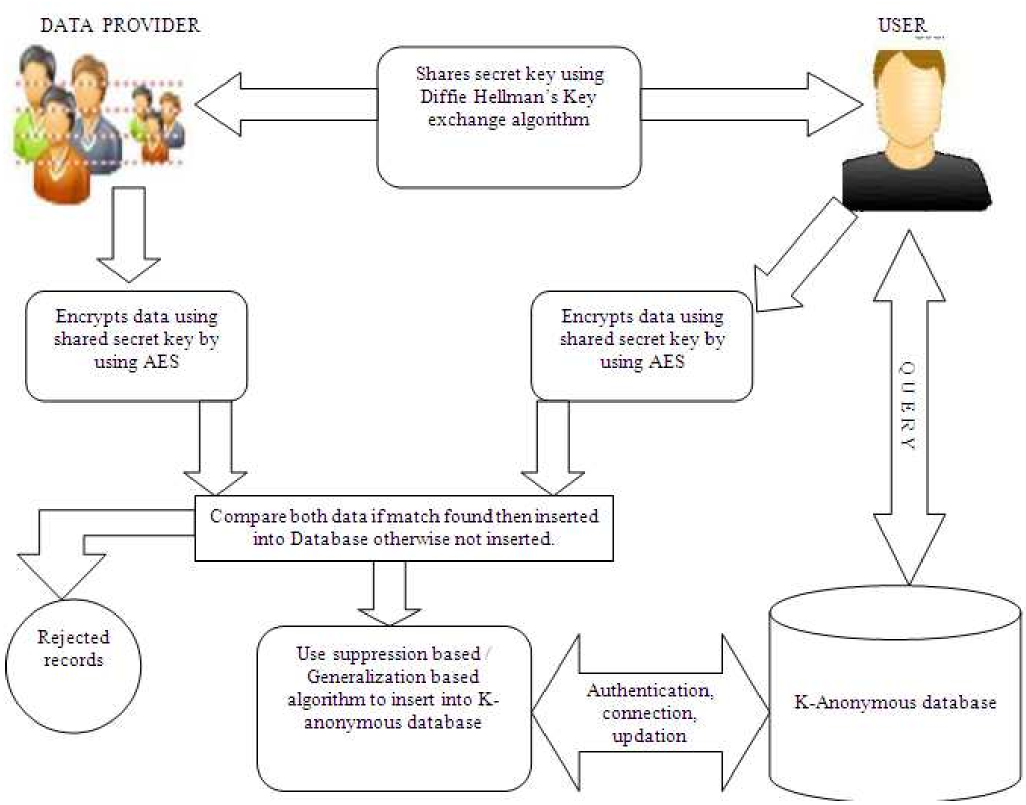

Figure 2 Overview of propose system

IV. WHY AES ALGORITHM

In this privacy preserving approach we use AES algorithm for improving the quality of overall system. The main reason for using AES is that it works under three approved key lengths: 128 bits, 192 bits, and 256 bits. An algorithm starts with a random number, in which the key and data encrypted with it are scrambled though four rounds of mathematical processes and make system stronger.

Another Diffie–Hellman key exchange algorithm is used for exchanging cryptographic keys. This algorithm allows two parties that have no prior knowledge of each other can share a shared key for communications by exchanging data over a public network. We are trying to solve the curse of dimensionality problem that are occurred in previous system and also improving the efficiency of system for this system gives response quickly with strong protection.

V. CONCLUSION

In this paper we observe various privacy preserving techniques, which fails on high dimensional data due to the curse of dimensionality. It causes too much information loss due to uniform distribution assumption. along with we have proposed two secure protocols which allows updating of K-anonymous database with maintaining its K-anonymity using AES technique. These protocols strongly protects, updates & maintain anonymity of database but not sufficient.

REFERENCES

[1] Alberto Trombetta, Wei Jiang, Member, IEEE, Elisa Bertino, Fellow, IEEE, and Lorenzo Bossi. July/August 2011. Privacy- Preserving Updates to Anonymous and Confidential Databases. IEEE Transactions On Dependable And Secure Computing, Vol. 8, No. 4.

[2] Elisa Bertino, Fellow, IEEE, and Ravi Sandhu, Fellow, IEEE January/march 2005. Database Security Concepts,Approaches, and Challenges. IEEE transactions on Dependable And Secure Computing, vol. 2, no. 1, january-march 2005

[3] Divya Sharma. April – 2012. Survey on Maintaining Privacy in Data Mining. International Journal of Engineering Research and Technology (IJERT). Vol. 1 Issue 2, April – 2012.

[4] Benjamin C. M.,Fung, Ke Wang, Rui Chen, Philip S. Yu. 2010.Privacy-Preserving Data Publishing: A Survey of Recent Developments. ACM Computing Surveys, Vol. 42, No. 4, Article 14.

[5] Gayatri Nayak , Swagatika Devi. March 2011. A Survey on Privacy Preserving Data Mining: Approaches and Techniques. International Journal of Engineering Science and Technology (IJEST), Vol. 3 No.3

[6] Mr. Mahesh T.Dhande1 Mrs. Neeta A.Nemade2 Research Scholar, Guide & Assi.Professor, S.S.G.B.C.O.E.T, Bhusawal, India S.S.G.B.C.O.E.T, Bhusawal, Indiajune2013. Performance Improvement of Privacy Preserving in K-anonymous Databases Using Advanced Encryption Standard Technique. Volume 3, Issue 6, June 2013 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering

[7] Benjamin C. M. Fung. 2007. Privacy-Preserving Data Publishing.Simon Fraser University

[8] Jaimin Marfatia1, Nainish Modi2, Niraj Lad3, Vaishali Patel4, Jignasa Patel5 Student, Dept. of Information Technology, Shri S’ad Vidhya Mandal Institute of Technology, Bharuch, India1,2,3 Professor, Dept. of Information Technology, Shri S’ad Vidhya Mandal Institute of Technology, Bharuch, India4,5.PRIVACY CONSERVING APPROACH to ANONYMOUS DATABASE.International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 1, March 2013.

[9] Kargupta H. Datta, S. Q. Wang and K. Sivakumar,”On the privacy preserving properties.