DISTRIBUTED PROTOCOL FOR PRIVACY - PRESERVING BY APPLYING PROPOSED K-ANONYMIZING APPROACH ON DISTRIBUTED DATA SETS

Padmapriya.G [1] *Hemalatha.M*[2]

ABSTRACT

Many Government agencies, business and organizations are willing to collect large amounts of data containing the sensitive information about individuals, such as income, diseases & symptoms also wants to release or share that data to other parties for research work. There is an increasing need for sharing data repositories containing personal information across multiple distributed and private databases. Concretely, given a query spanning multiple databases, query results should not contain individually identifiable information. So anonymization is done in privacy. Our works deal with privacy in database system. 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. It deals with privacy with anonymous database and on devising private update techniques to database systems that supports notions of anonymity.

Key Terms: - Privacy preservation,k-anonymization, confidentiality,suppression, Generalization.

I. INTRODUCTION

Today privacy or security has become crucial. So we mainly concentrate on privacy [2]. Privacy limits the access to individual's personal information. It deals with,authorized access. The collection of data usually called as the database, contains large bodies of information. To provide security to these databases is a big issue. Suppose person-A having his own k-anonymous database and person-B wants to insert a tuple. So, the problem is to check after inserting a tuple whether the database retains its k-anonymity or not. If allowing person-A to read the content of tuple directly, it breaks the privacy of person-B and on the other hand database confidentiality violated once person-B has access to the contents of the database, so privacy and confidentiality of the database are considered to be a major challenges. There are huge numbers of database containing numerous sensitive information (micro data) such as PIN, account number, income, diseases, etc., if all this information falls into wrong hands then it would be very danger-

Department of Computer science Karpagam University, Coimbatore. padmapriya.gokul@gmail.com

Department of Computer science Karpagam University, Coimbatore. csresearchhema@gmail.com

ous. If the database is not anonymous with respect to a tuple to be inserted, the insertion cannot be performed and updating is not possible. There are various anonymization techniques provides privacy protection which can be used such as data encryption, randomization and k-anonymity. Confidentiality means only authorized users can read the data. The extent of sensitive information about the citizens enrolled in the databases of government agencies and private organizations, such as census data, banks, student information and health care providers, has been increasing continuously in the past decades. It has recently become apparent that the information needs to be properly secured in case of transition and storage from unauthorized disclosure throughout the process of testing newly developed applications that utilize the databases. Now a day there is a great need of privacy of the users in the society. As we know that the use of computers is increasing in great amount, the requirement of privacy of each user and the confidentiality of the database is of the primary importance to the respective organization. There are huge numbers of databases stored in the system and by correlating these databases, private information of any specified user can be obtained. Hence, the database confidentiality and privacy of user is a big concern. In this paper, a method is proposed by which it is possible to maintain the privacy of each and every individual and simultaneously devise a method to preserve the confidentiality. Privacy is the data that can be securely shown to the valid owner without leaking the sensitive information from the database. Data confidentiality is the difficulty experienced by the third party to know any sensitive information stored in the database. Privacy is an essential issue in case of transpose sensitive information from one location to another location through internet. This issue hs arisen in different areas such as census, medical, financial transactions, governmental organizations and industries etc. Confidentiality can be termed as the preservation of information against unauthorized disclosure and limiting data access to authorized users. Data confidentiality is the nondisclosure of certain information except to authorize person [1], [2]. Microdata can be referred as data about individual, business, person or other entity and it can be collected by surveys, censuses or obtained from administrative records. It is stored in a table and each record (row) corresponds to one individual [3]. Microdata is an important issue in the public and the private sectors. Data anonymization enables transferring of information between the two organizations by converting text data into an unreadable form using encryption method. While the revealed data gives useful information to researchers which presents disclosure risk to the individuals whose data are present in the data [4], [5]. One of the methods used for protecting the privacy of user is to apply anonymization algorithms. *k*-anonymization technique is used for privacy preservation. A database is k-anonymous with respect to quasi-identifier attributes if there exist at least k transactions in the database having the same values according to the quasi-identifier attributes. The database is said to be k-anonymous where attributes are generalized or suppressed until each row is identical with at least k-1 other rows. This k-anonymity prevents distinct database linkages. It guarantees that the data released is accurate. If the value of k is large then better privacy is protected. It can also assure that individuals cannot be uniquely identified by linking attacks [14]. Overall a database is called k-anonymous Usually confidentiality can be achieved by using some cryptographic tools. Data Anonymization enables transferring information between twoorganizations, by converting text data into non-human readable form using encryption method [4].So the probability of linking a given datavalue to a specific individual is very small, and the individuals cannot be uniquely identified by linking attacks. This paper will give the solution. Two approaches can be used for Anonymization. Proposed system uses two manipulation techniques, suppression and generalization which are used to check that if new tuple is being inserted into the dataset it does not affect anonymity of databases. The proposed system uses commutative homomorphic encryption scheme to improve data privacy of the database and provides security of data by using AES algorithm.

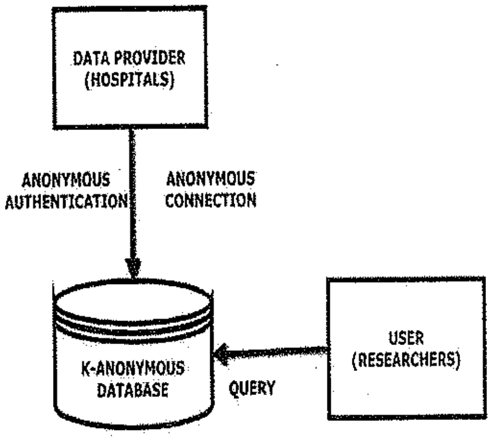

Figure 1: Anonymous Database

Figure: 1 shows, anonymous (sanitized) database system, which is used by the researchers. The dataproviders are medical facilities (Hospitals) through anonymous authentication and connection. Authenticationcan be done using user ID and password method. Anonymous connection can be done using Crowds or Onionrouting protocols. Anonymous updating is proposed in this paper. Crowds increase the privacy of webtransaction; the main idea is "blending into the crowd". That is, hiding one's action within the actions of manyothers. A user first joins a crowd of other users. The users request to a web server is first passed to a randomnumber of the crowd. That member can either submit the request directly to the end server or forward it toanother randomly chosen member. In the latter case the next member chooses to submit or forwardindependently. When the request is finally submitted, it is submitted by a random member, thus preventing theend server from identifying its true initiator. Even crowd members cannot identify the initiator of the request. Itis used for anonymous connection, it protects IP addresses and other sensitive information [7].Onion routingsupports private and anonymous connection/ communication over a public network. Onion routing is a flexiblecommunication infrastructurethat is resistant to both eavesdropping and traffic analysis. It is a bi-directional,near real- time and can be used for both connection oriented and connection less traffic. When a packet isreceived by the first onion router, it is encrypted once for each additional router it will pass through. Eachsubsequent Onion router unwraps one layer of encryption until the message reaches its destination as plain text[6]. Developed for privacy conservation, but one well known technique is K- anonymization. Anonymizationtechniques enables transferring between the two organizations by converting text into humanreadable form using the encryption method [1].Anonymization is about preserving identifying or privateinformation using encryption. Nonanonymized version of data deleted from the sender side after it is being sent tothe receiver side. This is one of the important concepts of this technique [2].Such technique provides security bymodifying data in such a way that it gives the same results for more than two tuples. When a provider wants toinsert a tuple into database, then it gives birth to two problems concerning both the individual's privacy as wellas confidentiality of databases.

i) Is updated database still privacy preserving.

ii) Does the database owner needs toknow the data to be inserted.

II. Literature Survey

Trombetta et al. [10] suggested method deals with problems regarding the users without revealing the contents of tuplesand database, how to protect data integrity by implementing the anonymity of database and if the anonymity is authorizedthen there is a concern of updating the data. It deals with algorithms for database anonymization.In paper [22] Trombetta et al. checked whether the database adjoined with the tuple remains k-anonymous or itsanonymity is violated. But these protocols do not support generalization-based updates which is the main approachadopted for data anonymization. Therefore if the database is not anonymous with respect to a tuple to be inserted, theinsertion cannot be performed and so updation of database is also not possible.Wortmann et al. [11] deals with algorithms for Database anonymization. Idea of protecting database through datasuppression or data perturbation has been extensively investigated in 1989. Problem of security-control methods is notsuitable when multiple attribute databases are used [4] Yehuda Lindell et al. [12] proposed a Secure Multi-party Computation (SMC) technique in 2009. SMC technique encrypts the data sets, while still allowing data mining tasks. It is not supposed to disclose any new information other than the final result of the computation to a participating party. These techniques are usually based on cryptographicprotocols and are applied to distributed data sets. It suffers from the limitations that this method becomes complicated when more than two parties are involved and it does not address the question of whether the disclosure of the final data mining result may violate the privacy of individual records.Fung et al. [15] proposed Top-down Specialization (TDS) approach for handling both categorical and continuous attributes of data. This approach focuses on exploitation of the fact that data usually contains redundant structures for classification. They claimed that TDS is much more effective than genetic algorithm and it scales well with large data sets and complex anonymity requirements. But effectiveness of TDS decreases in handling.

LeFevre et al. [16] presented a greedy top-down specialization algorithm for finding a minimal k-anonymization in the case of the multidimensional generalization scheme. The greedy algorithm used for multidimensional partitioning performs better than the other optimal algorithm but it is expensive algorithm.Top-down refinement method for k-anonymization solution for classification was suggested by Benjamin & Fung et al in 2007 [17]. TDR is capable of suppressing a categorical attribute with no taxonomy tree. In TDR each refinement increases the information and decreases the anonymity since records with specific values are more distinguishable. They use a single dimension recoding. . This "over-suppression" reduces the quality of the anonymous datasets.Arik Friedman et al. [18] proposed a decision tree induction algorithm is guaranteed to maintain k-anonymity in 2008. It differs from other methods such as TDS and TDR by letting the data owners share with each other the classification models extracted from their own private datasets, rather than to let the data owners publish any of their own private datasets. Therefore the output of kADET is an anonymous decision tree rather than an anonymous dataset.In 2006, S. Zhong et al. [19] prove that it does not reveal any additional information and protocols improve the privacy of k-anonymization by maintaining end-to-end privacy from the original customer data to the final k-anonymous results.G. Aggarwal et al. [20] present an approach in 2005 in which it has the problem of releasing tables from a relational database containing personal records is considered and how it can be resolved, ensuring personal privacy and also maintaining integrity.E. Bertino et al. [21] present an approach privacy preserving incremental data dissemination in 2009. This paper is to identify and prevent cross- version inferences so that an increasing dataset can be incrementally disseminated without compromising the imposed privacy requirement. Disadvantage is updation of record is not possible and also they not used any cryptography method so does not guarantee about privacy.There are various techniques like data perturbation, query processing, anonymizing tables, TDS, TDR, kADET and SMC provides confidentiality and privacy to anonymous database. But the main problem is that, a result does not deal with the updations in the database, which results in the privacy break. Therefore none of the work resulted in achieving the privacy and confidentiality

III. Methods For Privacy Preserving

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

A) Cryptographic approach

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

Advantages

1. It is a 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 a few parties involved.

B) Bucketization

Bucketization removes the identifiers from the data and also partitions tuples into buckets. Buckets containthe subset of tuples. Generalization transforms the QI values in each bucket into "less specific, but semanticallyconsistent" values. So that tuples of the same bucket cannot be distinguished by their QI values. In bucketizationseparates the SAs from the QIs but randomly permuting the SA values in eachBucket.

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 easilyimplemented because the noise added to the given record is independent from the other records. The amount ofnoise is large enough to smear original values, so individual record cannot be recovered. The randomizationmethod is simple as compare to other methods because it does not require knowledge of other records. Largerandomization increases the uncertainty and user's personal privacy. They claim that approaches may loseinformation as well as not provide privacy by introducing random noise to the data by using random matrixproperties, [13]. It successfully separates the data from the random noise and subsequently discloses the originaldata.

RandomizeAdvantages

1. The randomization method is very simple method and which can be easily applied when we collect thedata.
2. It is support protecting 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 a very slow technique because when data collector collects the data from data provider the dataprovider 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 privateinformation. There are many ways to perform data anonymization basically this method uses k-anonymizationapproach 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 andgender. There are k people meet the requirement [1][3][4].

IV. Proposed System Of K-Anonymization

Anonymization is about preserving identifying or private information using encryption. The main purpose isto protect sensitive information. In a k-anonymous dataset, if any identifying information is found in the originaldataset with k tuples then first we identifies quasi-identifiers i.e. the tuple that clearly distinguish the given tuplein 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 atleast K rows & the union of these subset id the entire table, then we are replacing each records with '*'.Insuppression based algorithm we are using diffie Hellman Key exchange algorithm [10] to generate privatesecure key. Then we are applying the AES (Advance Encryption Standards) algorithm [10] to encrypt &decryptdata by using the key generated by diffie Hellman key exchange algorithm. In this approach we are dealing withencrypted data not with original data. When user enters their information, then we encrypt it by using AESalgorithm simultaneously we also encrypt the data in table using same algorithm. If information inserted by usermatches 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 valueIV. 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.To understand the concept of data anonymization considers a simple example of the medical patient. Theinformation about a single patient is stored in a single line, i.e. tuple, and database is store confidentially at theserver. The users may be medical researchers who have the access to the DB. Since DB is anonymous, one mainconcern is to protect the privacy of patients. Such task is guaranteed through the use of anonymization. if thedatabase DB is anonymous, it is not possible to identify the patients record.

There are two mostly used anonymization techniques as follows

5.1 Suppression based K-anonymization

The basic concept of suppression based algorithm is to mask some attributes by special value *,In this algorithmt stands for tuple which is to be inserted by data provider & T stands for the Anonymous Database. QI stands forQuasi-Identifier which consists of set of attributes that can be used with certain external information to identifya 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 versionof 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 totable T. Otherwise, when inserted to T, t breaks k-anonymity.

5.2 Generalization Based K-anonymization

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

The protocol works as follows:

Step 1: User X randomly chooses a ă å *Tw* (Witness Set).

Step 2: User X computes ã = GetSpec (ă).

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

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 represents the actual information in the original Dataset, after applying the suppression based algorithmover the original dataset the original dataset is anonymized& displays anonymized records. The "Data Mining" point can be generalized to more specificvalue with "Database System".So, by applying this concept & replacing the remaining values in table with a more general value the originaldataset is anonymized using generalized method & finally when T is K-anonymous, we can delete duplicatetuples, & 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 a more possiblecombination related to original datasets.

Demerits:

i) The main problems with generalization are it fails on high-dimensional data due to the curse ofdimensionality 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]

V. Algorithm

STEP 1: X encrypt the tuple T, and send it to Y.

STEP 2: Y can decrypt tuple T and then suppress the personal identifiers in the tuple.

STEP 3: After the suppression checks the nonsuppresed attributes in the tuple T and loaded tuples.

STEP 4: If any match found, insertion can be performed and send a status message "INSERTED".

STEP 5: If no match found, discard the tuple and send the status message "IGNORE".

VI. Conclusion

In this paper we observe various privacy preserving techniques, which fails on high dimensional data due tothe curse of dimensionality. It causes too much information loss due to the uniform distribution assumption. Alongwith we have proposed two secure protocols which allows updating of K-anonymous database with maintainingits K-anonymity using AES technique. These protocols strongly protects, updates & maintain anonymity ofdatabase but not sufficient.we have proposed secure protocol to check that if new tuple is being inserted to the database, itdoes not affect anonymity of database. It means when new tuple gets introduced, k-anonymous database retainsits anonymity. Database updates have been carried out properly using proposed protocol. Thus, by making such kanonymityin a table that makes unauthorized user too difficult to identify the record.

REFERENCES

[1] G Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy,D. Thomas, and A. Zhu,"Anonymizing Tables," Proc. Int'l Conf. Database Theory (ICDT), 2005.

[2] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee, "Towards Privacy in Public Databases,"Proc. Theory of Cryptography Conf. (TCC), 2005.

[3] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertainty, Fuzziness andKnowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.

[4] Privacy-Preserving Updates to Anonymous and Confidential Databases, Alberto Trombetta, Wei Jiang,Elisa Bertino and Lorenzo Bossi, Department of Computer Science and Communication, University ofInsubria, Italy.

[5] Generalization Based Approach to Confidential Database Updates, Neha Gosai, S H Patil, Department ofComputer Science, Pune, Maharashtra, 2012

[6] Murdoch Steven J, Danezis G.low-cost traffic analysis[7] TOR In: IEEE symposium on security and privacy May 2005.

[7] Brier, S. 1997. How to keep your privacy: Battle lines get clearer. The New York Times, January 13,1997.

[8] G. Aggarwal, N. Mishra and B. Pinkas. Secure Computation of the k-th Ranked Element. InEUROCRYPT 2004, Springer-Verlag (LNCS 3027), pages 40{55, 2004}.

[9] J. Li, N. Li, W. Wins borough. Policy-hiding access control in open environment. In Proc of ACM Conf.on Computer and Communications Security (CCS), Alexandria, Virginia, 2005.

[10] N.R. Adam and J.C. Worthmann, "Security-Control Methods for Statistical Databases: A ComparativeStudy," ACM Computing Surveys (CSUR), vol. 21, no. 4, pp. 515-556, 1989.

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

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

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

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

[15] GayatriNayak, 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

Author's Biography

Padmapriya.G, received the first degree in B.C.A from Bharathiyar University in 2004, Tamilnadu, India. She obtained her master degree in Computer Communication from Bharathiyar university and She obtained her master degree in Computer Applications from Bharathyiar University in 2012, Tamilnadu, India. She is currently pursuing her Ph.D. degree Under the guidance of Dr. M.Hemalatha, Head, Dept of Software Systems, Karpagam University, Tamilnadu, India.

Dr. M.Hemalatha, completed M.Sc., M.C.A., M. Phil., Ph.D (Ph.D, Mother Terasa women's University, Kodaikanal). She is Professor & Head and guiding Ph.D Scholars in Department of Computer Science in Karpagam University, Coimbatore. Twelve years of experience in teaching and published more than hundred papers in International Journals and also presented more than eighty papers in various national and international conferences. Area of research is Data Mining, Software Engineering, Bioinformatics, and Neural Network. She is a Reviewer in several teaching and published more than hundred papers in International Journals and also presented more than eighty papers in various national and international conferences. Area of research is Data Mining, Software Engineering, Bioinformatics, and Neural Network. She is a Reviewer in several National and International Journals.