ISSN:0975-9646

Privacy Preserving Updates to Sliced Anonymous Data Bases

Asha Thomas
Dept. of Computer Science & Engg.
Sree Buddha College of Engg.
Alappuzha, Kerala, INDIA

Adarsh Sunil
Dept. of Computer Science & Engg.
Sree Buddha College of Engg.
Alappuzha, Kerala, INDIA

Abstract—**Privacy is the main concern in the present world. Privacy is becoming an increasingly important issue in many data mining applications in various fields like medical research, intelligence agencies, hospital records maintenance etc. Suppose there exsists an anonymous database (e.g. containing medical records) then the objectives include how to still preserve the privacy while updates are being made into the current anonymous database by preserving the privacy of the user and confidentiality of the database. It is required to ensure that the database is still anonymous after the update. In this paper the database is made anonymous by slicing approach and then updates are made to the sliced database. Slicing overcomes the disadvantages of k-anonymity.**

Keywords—slicing, confidentiality, generalization, k-anonymity, privacy, SMC.

I. Introduction

In today's world database is a valuable asset for many different applications, so security becomes critical. Bank information, medical research database – all these information can be dangerous if it fall into wrong hands. There is big concern of privacy.

Data in the databases has its own relevant value. For example, medical data collected by over the history of patients over years is an invaluable asset, which needs to be secured. Nowadays, privacy has become common problem in the information systems. For example, a hospital may have record of all the patients with various diseases critical and non-critical. If the hospital wishes to reveal the data to any pharmaceutical company or online market services, it should not be able to infer with particularity of patients with those diseases. It can give as a statistical view or just the superficial information such that privacy is not detained.

Detailed person-specific data in its original form often contains sensitive information about individuals, and publishing such data immediately violates individual privacy. The current practice primarily relies on policies and guidelines to restrict the types of publishable data and on agreements on the use and storage of sensitive data. The limitation of this approach is that it either distorts data excessively or requires a trulevel that is impractically high in many data-sharing scenarios. For example, contracts and agreements cannot guarantee that sensitive data will not be carelessly misplaced and end up in the wrong hands. Today, with the increased storage of personal data of musers on Internet, problem of privacy preserving data mining has become very crucial. Organizations and different agencies very often need to publish sensitive micro data, like medical data or census data or student information data for research purpose or for developers to develop applications.

Decision makers and trend analyzer also need such type of micro data. The term micro data can be referred to as data about an individual, person, household, business or other entity. Micro data may be data collected by surveys, censuses or obtained from administrative records. This must be done in such a way that the confidentiality of the information provided by respondents is preserved. With these recent changes in trends, sensitive data is now easily available for malicious use. The main concern is that sensitive information should not be disclosed or misinterpreted. Confidentiality can be termed as limiting data access and disclosure to authorized users and preventing access by or disclosure to unauthorized ones.

For example, consider the hospital data pertaining patients information The data contains attribute values which can uniquely identify an individual(pincode, nationality, age) or/and (name) and sensitive information corresponding to individuals (medical condition, salary, location). The organization wants to wants to publish in such a way that information remains practically useful and also identity of an individual is not compromised. With the approaches of Anonymization Techniques various classified fields based on their importance can be protected.

Privacy is the right of individual person to keep their information secretly hidden from others. Data confidentiality is the non disclosure of certain information except to authorized person. Confidentiality relates to data and privacy relates to person. The term anonymized or anonymization means identifying information is removed from the original data to protect personal or private information. This technique protects privacy of original data by modification. Anonymisation modifies the original table so that the sensitive values are not inferred.This helps in preserving the privacy of the user.

II. Related Works

The following section presents summaries on various approaches that exist for preserving the privacy in data bases. Certain approaches mentioned have disadvantages. The k-anonymity approach ensures privacy even after updates are being made to anonymous databases but that approach too has drawbacks and thus comes the concept of slicing as an anonymisation approach in databases.

A. Randomisation Approach

In paper [1], a randomization technique was proposed to preserve the privacy.In this approach, randomization noise is added to the data so that the behavior of the original data is masked. The randomization method provides effective way of preventing 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. If the the set of data records is denoted by x1,x2,….xn.The noise component is drawn from the probability distribution function f(Y).These noise components are denoted by y1,y2…..yn.Thus the new set of records obtained is z1,z2….zn which is the summation of the x's and y's.It is possible to obtain the original records by subtracting y from z as n instances of records are being known. This proves to be a major disadvantage of this approach.

B.Secure Multi Party computation Approach

In [2], Secure multi-party computation (SMC) approach is proposed. With SMC, several parties can jointly perform some global computation on their private data without any loss of data security/privacy. Consider a patient has been ill for last 5 years. He has taken treatment from several doctors and sometimes he has been hospitalized too. If it is needed to calculate the complete recovery time of that patient, it will be the joint sum of durations for which patient took treatment from each individual doctors and the duration for which he was hospitalized. Each doctor and hospitals maintain their patient's database. Now here joint computation is involved and this computation only provides recovery duration without revealing other information of any doctor's clinic or hospital databases. This technique is not efficient as the method needs information from individual data bases. Hence computations cannot be performed securely.

C.Private Information Retrieval Approach

In [3],PIR approach is proposed.PIR means private information retrieval.PIR provides a means to retrieve data from a database without revealing any information about which item is retrieved. In its simplest form, the database stores an n-bit string X, organized as r data blocks, each of size b bits. The user's private input or query is an index i ∈ {1, ..., r} representing the i th data block. A trivial solution for PIR is for the database to send all r blocks to the user and have the user select the block of interest at index i (i.e., Xi), but the complexity involved is high which is considered to be a disadvantage of PIR.

D.K Anonymisation Approach

In [4],authors proposed k-anonymity technique. The k anonymity model ensures that each record in the table is identical to at least (k-1) other records with respect to the privacy-related features. Therefore, no privacy related information can be inferred from the k-anonymity protected table during a data mining process. In the k anonymity model, the quasi-identifier feature set consists of features in a table that potentially reveals private information, possibly by joining with other tables. K-anonymity involves two approaches generalization and suppression.

Generalisation involves generalizing the quasi identifier fields and suppression involves suppressing the values in those fields.Generalisation approach is the most commonly used approach and it has certain disadvantages like Generalisation approach suffers from the curse of dealing with high dimensional data,reduces the data utility of the generalised data and Corelation between different attributes are lost. Thus the concept of slicing is introduced.

III. PROBLEM STATEMENT

Problem could be stated as: How updates could be made to sliced anonymous databases thus ensuring the privacy of the user and the confidentiality of the database.

The problem could be explained in more detail as if an anonymous database is owned by a database owner and suppose a user wants to insert certain tuples or update an exsisting tuple then in such a case the update should be made without revealing the the contents to the database owner and the contents of the database is not revealed to the user. Disclosing the database contents looses confidentiality and disclosing the tuple contents breaks the privacy of the user.

The major objectives of the thesis are

*1)*To devise a privacy updating technique to database systems that support notions of anonymity other than k-anonymity.
Disadvantage of k-anonymity

- ✓ Reduces the data utility of the generalized data.
- ✓ Correlation between different attributes are lost.

  Sicing anonymisation technique overcomes the above disadvantages.

*2)*To devise techniques to deal with the tuples that fail to satisfy the anonymity.

*3)*To ensure more privacy and confidentiality of the database by introducing anonymity.

The first and foremost objective to be met is the base work of this project. Privacy updating technique other than k anonymity introduced here is Slicing. Slicing preserves utility because it groups highly correlated attributes together, and preserves the correlations between such attributes. Slicing partitions the data set both vertically and horizontally.

IV METHODOLOGY

Table containing the details of the patients which includes the age,sex,pincode and sensitive information disease is given as the input for anonymisation using slicing.The project could be splitted into three modules.

1) Creation of an anonymous database using Slicing.

The creation of an anonymous database using the technique of slicing involves three different phases as follows:

i. Attribute partitioning.

This is also known as vertical partitioning. This step partitions attributes so that highly correlated attributes are in the same column.e.g (age,sex).

ii. Column generalization.

This step involves generalizing the values in the columns so that the database satisfies the anonymity requirement.

iii. Tuple partitioning

This is also known as horizontal partitioning. This step involves partitioning the tuples into buckets.

iv. Random permutation of values

The values in the partitioned columns are randomly permuted so that it is not possible to discover the sensitive values as the values in each tuple gets changed after permuting the values.

2) Updating the sliced anonymous database

After anonymous database gets created insertions of new tuples or updation of exsisting tuples into the sliced anonymous database can be done.this is made possible by performing Re-slicing.Re-slicing performs the reverse steps of the slicing algorithm.It first performs reverse of tuple partitioning where buckets are removed.Then reverse of attribute partitioning is performed where the coloumns are splitted in the original form. Thus the original table is obtained.insertion and updation is made into this original table.After updation slicing is again performed.In the case of a medical application updates could be made by the doctor of their repective patients only all other patient's details remains in the anonymised form. Patients could do updates only to their personal records and the medical researchers could only access the anonymised table.

3)Handling of pending tuple set

In some cases the tuples being updated or that is being inserted may fail to satisfy the anonymity requirement. Then such tuples may fall into the pending tuple set.this module deals with the re insertion of such tuples after verifying the anonymity by using the concept of time stamping. Updates could be performed by comparing the tuples that failed to get inserted into the database with their original timestamp values.

IV RESULT

Anonymous table gets created using the concept of slicing. The original table gets modified where the original attributes are partitioned into columns of highly correlated attributes and the tuples are partitioned into buckets and the values in the columns are randomly permuted. This preserves the privacy and thus breaks the issue of disclosing the sensitive values.

Updates could be made to such a sliced anonymous data base by the process of re-slicing. By doing re-slicing all the process taken place during slicing gets reversed. Thus updation is made to the original table and then again converted back to the sliced form. Thus the anonymity is preserved.

V CONCLUSION

The objective of devising a private update technique other than k-anonymity could be obtained by using the concept of slicing. Slicing preserves the privacy by horizontal and vertical partitioning of the tuples.Updates could be made to the sliced anonymous tables using the concept of Re-slicing. Slicing overcomes the limitations of generalization and preserves better utility while protecting against privacy threats.The concept of anonymisation ensures that only authorized users can view the sensitive information and to other users the database appears in the anonymised form.

REFERENCES

[1].Kargupta H. Datta, S. Q. Wang and K. Sivakumar ,"On the privacy preserving properties of random perturbation techniques"IEEEICDM,2003

[2].Dr. Durgesh Kumar Mishra1, Purnima Trivedi2,Samiksha Shukla3," A Glance at Secure Multiparty Computation for Privacy Preserving Data Mining" International Journal on Computer Science and Engineering Vol.1(3), 2009.

[3].Femi Olumofin and Ian Goldberg"Privacy-preserving Queries over Relational Databases", In Network and Distributed Systems Security Symposium, 2010.

[4]. L. Sweeney, "k-anonymity: a model for protecting privacy." International Journal on Uncertain. Fuuiness KnowL-Based Syst**.,** vol. 10, no. 5, pp. 557-570, 2002.

[5]. L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," International Journal on Uncertainty,Fuzziness and Knowledge-based Systems**,** vol. 10, no. 5, 2002.

[6].Ashwin Machanavajjhala,Johannes Gehrke,Daniel Kifer,"ℓ-Diversity: Privacy Beyond k-Anonymity", In IEEE Transactions on Knowledge and Data Engineering, 2009.

[7].Tiancheng Li, Ninghui Li, Jian Zhang,"Slicing: A New Approach for Privacy Preserving Data Publishing", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 3, MARCH 2012

[8]**.**Kavitha,"Suppression and Generalization – Based Privacy Preserving Updates to Confidential Databases", IOSR Journal of Computer Engineering (IOSR-JCE),Volume 10, Issue 1 (Mar. - Apr. 2013), PP 51-54.

[9] A. Trombetta, E. Bertino. Private updates to anonymous databases. In Proc. Int'l Conf. on Data Engineering (ICDE), Atlanta, Georgia, US, 2006.