International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 20 Issue 2 – FEBRUARY 2016.

ENHANCED PRIVACY IN INCREMENTAL DATA ANONYMIZATION WITH PRIVACY PRESERVING ACCESS CONTROL MECHANISM

P.Sivakamasundari[1], S.Lavanya[2]

[1]*Assistant Professor, Dept of CSE, Adhiparasakthi Engineering College, Melmaruvathur*
[2]*PG Scholar, Dept of CSE, Adhiparasakthi Engineering College, Melmaruvathur*

sivapreethi18@gmail.com
lavanya2692011@gmail.com

***Abstract*--- Access control mechanisms are used to ensure that sensitive information is available to authorized users only. When Privacy Protection Mechanism (PPM) is not used, authorized users can misuse the sensitive information and the privacy of the consumer is compromised. Privacy requirement is satisfied by PPM which uses suppression and generalization approaches to anonymize the relational data. K-anonymity or l-diversity is used to anonymize and satisfy privacy requirement. However, privacy is obtained by the precision of the authorized information. The anonymity technique can be used with an access control mechanism to ensure both security and privacy of the sensitive information. In this paper Role based access control is assumed. The access control policies define selection predicates to roles. Then we use the concept of imprecision bound for all permission to define a threshold on the amount of imprecision that can be tolerated. So the proposed approach reduces the imprecision for each selection predicate. Anonymization is carried out only for the static relational table in the existing papers. In this paper privacy preserving access control mechanism is applied to the incremental data.**

I. INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analysing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analysing data. It allows users to analyse data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD).

Data storage became easier as the availability of large amounts of computing power at low cost i.e. the cost of processing power and storage is falling, made data cheap. There was also the introduction of new machine learning methods for knowledge representation based on logic programming etc. in addition to traditional statistical analysis of data. The new methods tend to be computationally intensive hence a demand for more processing power. Data mining analysis tends to work from the data up and the best techniques are those developed with an orientation towards large volumes of data, making use of as much of the collected data as possible to arrive at reliable conclusions and decisions. The analysis process starts with a set of data, uses a methodology to develop an optimal representation of the structure of the data during which time knowledge is acquired. Once knowledge has been acquired this can be extended to larger sets of data working on the assumption that the larger data set has a structure similar to the sample data. Again this is analogous to a mining operation where large amounts of low grade materials are sifted through in order to find something of value. Applications of data mining are financial data analysis, biological data analysis, retail industry, telecommunication industry and other scientific applications.

II. RELATED WORKS

A. Generalization Algorithm

In this method, individual values of attributes are replaced by with a broader category. For example, the value '19' of the attribute 'Age' may be replaced by ' $\leq$ 20', the value '23' by '20 < Age $\leq$ 30' , etc. The attributes available to an adversary are called "quasi-identifiers". Each "quasi-identifier" tuple occurs in at least k records for a dataset with k-anonymity. Data Generalization is the process of creating successive layers of summary data in an evolutional database. It is a process of zooming out to get a broader view of a problem. Having data from several sources greatly helps in the overall business intelligence system of a company.

B. Suppression Algorithm

In this method, certain values of the attributes are replaced by an asterisk '*'. All or some values of a column may be replaced by '*'. In the anonymized table below, we have replaced all the values in the 'Name' attribute and all the values in the 'Religion' attribute have been replaced by '*'.

Protecting individual data in disclosed databases is essential. Data anonymization strategies can produce table ambiguation by suppression of selected cells. Using table ambiguation, different degrees of anonymization can be achieved, depending on the number of individuals that a particular case must become indistinguishable from. This number defines the level of anonymization. Anonymization by cell suppression does not necessarily prevent inferences from being made from the disclosed data. Preventing inferences may be important to preserve confidentiality.

III. SYSTEM ARCHITECTURE

The architecture involves the processing of collecting dataset and anonymzing the dataset for privacy purpose. Datasets are available in various websites. So we can easily download the dataset for our project.

Architecture model

Datasets are not in a proper format. So we cannot be able to do any process on the raw data. So we need to pre-process the table. Preprocessing is the process of forming table from the dataset. After that we need use the anonymization concept for privacy. Before anonymization we need to identify the sensitive attributes. Sensitive attributes are attributes by using that we can easily identify the personal details. For hiding or changing the originality of the sensitive attributes we need to do the anonymization.

After anonymization process we need to add role based access control mechanism. We can create anonymization table with bound by using top down heuristics algorithm.

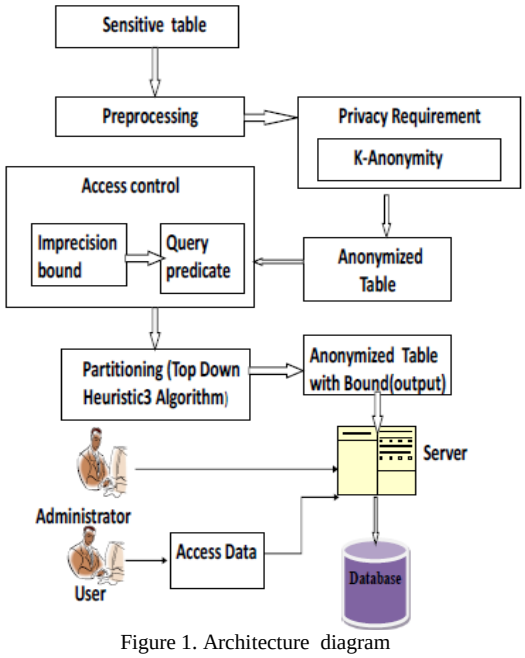

Figure 1. Architecture diagram

IV. SYSTEM MODULES

A. Preprocessing

Data preprocessing describes any type of processing performed on raw data to prepare it for *another* processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

In this module, Dataset is taken from file which is downloaded from web. For example UIMechanism. Dataset contains raw data. It will not be in proper format. We cannot use that raw data for anonymization. So we have to make it into proper format. Dataset contains data in the text file is converted in to the table form for further processing. Forming table from the selected Dataset is known as preprocessing. After that anonymization technique is applied for the Dataset.

B. Cluster Formation

A cluster is a subset of objects which are similar. Clustering is a process of partitioning a set of data into a set of meaningful sub-classes, called clusters. Help users understand the natural grouping or structure in a data set. Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.

After preprocessing we can get a table with lots of data. From that table we do the anonymization process. In that we have to select related attribute to form cluster. Clustering is the process of partitioning. By using this clustering concept we can easily anonymize data. If we didn't select related attribute, it will be very tough to search and find data.

C. Anonymization

Data anonymization is a type of information sanitization whose intent is privacy protection. It is the process of either encrypting or removing personally identifiable information from Datasets, so that the people whom the data describe remain anonymous. By clustering process we can partition the large database.

Now we have to anonymize the data. Anonymization is the process of converting a text in a range of value or into a non-readable symbol form. After clustering process completed we need to anonymize data using algorithms. Anonymization process will improve efficiency of data. Because of anonymization we can also save time. We can use methods like Generalization and Suppression to anonymize data.

D. Collecting Anonymized Data

Because of clustering we cannot get all anonymized data together. We will get cluster by cluster after anonymization. Now we have to gather all data after anonymization. So we can store the total anonymized data in the server. Then we can release this anonymized dataset for further use. So we can efficiently use the dataset with privacy.

V. METHODOLOGY

A. String searching algorithm

String-matching is a very important subject in the wider domain of text processing. String-matching algorithms are basic components used in implementations of practical software existing under most operating systems. By using this string pattern matching algorithm we can easily do the preprocessing step. The downloaded

Dataset is not in a clear form. To make those data in a clear form or in a table form we need to do preprocessing.

Moreover, they emphasize programming methods that serve as paradigms in other fields of computer science (system or software design). Finally, they also play an important role in theoretical computer science by providing challenging problems. Although data are memorized in various ways; text remains the main form to exchange information. This is particularly evident in literature or linguistics where data are composed of huge corpus and dictionaries. This is the reason why algorithms should be efficient even if the speed and capacity of storage of computers increase regularly.

B. Anonymization algorithm

Data anonymization is a type of information whose intent is privacy protection. It is the process of either encrypting or removing personally identifiable information from Datasets, so that the people whom the data describe remain anonymous. The Privacy Technology Focus Group defines it as "technology that converts clear text data into a nonhuman readable and irreversible form.

After preprocessing we can get clear data from the dataset. Our main aim of this project is data anonymization. So we can give privacy to the data owner. Sensitive attributes are not revealed to the user. Anonymization algorithm is used to anonymize the data by using some symbols. After anonymization user will not understand all the data. Particularly they cannot understand the sensitive attributes. Sensitive attributes will be changed into non readable symbols.

VI. CONCLUSION

Global recoding may recode more than needed, whereas local recoding complicates data analysis by mixing together values corresponding to different levels of generalization. Also, recoding produces a greater loss of granularity of the data, is more affected by outliers, and changes numerical values to ranges. Regarding local suppression, it complicates data analysis with missing values and is not obvious to combine with recoding in order to decrease the amount of generalization. Microaggregation is free from all the above downsides. We have proposed and evaluated three different microaggregation based algorithms to generate k- anonymous t-close data sets. The first one is a simple merging step that can be run after any microaggregation algorithm. The other two algorithms, k-anonymity-first and t-closeness-first, take the t-closeness requirement into account at the moment of cluster formation. The t-closeness-first algorithm considers t-closeness earliest and provides the best results: smallest average cluster size, smallest SSE for a given level of t-closeness, and shortest run time. Thus, considering the t-closeness requirement from the very beginning turns out to be the best option.

REFERENCES

[1]J. Domingo-Ferrer and J. Soria-Comas. From t-closeness to differ-ential privacy and vice versa in data anonymization. Knowledge- Based Systems, 74:151–158, 2015.

[2]L. Sweeney. k-anonymity: a model for protecting privacy. Interna- tional Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):557–570, 2002.

[3]J. Soria-Comas and J. Domingo-Ferrer. Differential privacy via t-closeness in data publishing. In Proceedings of the 11th Annual International Conference on Privacy, Security and Trust (PST 2013), pp. 27–35, 2013.

[4]J. Li, R.C.-W. Wong, A.W.-C. Fu, and J. Pei. Anonymization by local recoding in data with attribute hierarchical taxonomies. IEEE Transactions on Knowledge and Data Engineering, 20(9):1181–1194, 2008.

[5] J. Domingo-Ferrer and V. Torra. Ordinal, continuous and het-erogeneous k-anonymity
through microaggregation. Data Min. Knowl. Discov., 11(2):195–212, 2005.

[6] J. Cao, P. Karras, P. Kalnis, and K.-L. Tan. SABRE: a Sensi- tive Attribute Bucketization and REdistribution framework for t-closeness. The VLDB Journal, 20(1):59-81, 2011.

[7]J. Domingo-Ferrer and V. Torra. A quantitative comparison of disclosure control methods for microdata. In Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies (eds. L. Zayatz, P. Doyle, J. Theeuwes and J. Lane), pp. 111–134, Amsterdam, 2001. North Holland.

[8]J. Soria-Comas, J. Domingo-Ferrer, D. S´anchez and S. Mart´ınez. Enhancing data utility in differential privacy via microaggregation-based k-anonymity. VLDB Journal 23(5):771– 794, 2014.

Authors Profile

P.Sivakamasundari received the **B.E.** degree in Computer science and Engineering from Adhiparasakthi Engineering College, Anna University, Chennai, India in 2005. Received **M.E.** degree in Computer Science and Engineering in Sri Krishna Engineering College, Anna University, Chennai, India in 2011. Her research interest includes Networks.

S. Lavanya received the **B.Tech.** degree in Information Technology from IFET Engineering College, Anna University, Chennai, India, in 2007. Currently doing **M.E.** in Computer Science and Engineering in Adhiparasakthi Engineering College, Anna University, Chennai, India. Her research interest includes Data mining, Access control mechanism and Data anonymization.