A STUCTURE TO USE DIFFERENT ANONYMIZATION TECHNIQUES FOR PRIVACY PRESERVING DATA PBLISHING

Mr. Lahare Prasad A.
Department of Computer Engineering, Amrutvahini College of Engineering,
Sangamner, SPPU Pune, Maharashtra, India.

Prof. M.A. Wakchaure
Assistant Professor, Department of Computer Engineering,
Amrutvahini College of Engineering, Sangamner, SPPU Pune, Maharashtra, India.

ABSTRACT

In this paper we show the framework of Slicing with Tuple grouping algorithm which partitioned the data both horizontally and vertically. It provides better information utility than generalization and Bucketization. Privacy preservation is important for publishing the personal information. Generally personal information records will violate the privacy. So many more techniques have been introduced for privacy preservation. Many anonymization techniques like generalization and bucketization have been designed and developed for privacy preservation. But they have some disadvantages. In this survey paper, we present technique called as slicing, which partitions the data both horizontally and vertically. We experimentally show that how slicing preserves better data information utility than generalization and handle in high dimensional data and protect from membership disclosure.

INTRODUCTION

In past few years, due to increase in ability to store personal information about local and global users and the increasing data mining algorithms to secure and full fill this information the problem of privacy-preserving data mining has become more important today. A number of anonymization techniques have been introduced in order to perform privacy-preserving data mining. Most of existing work is performed in the following manner: Several institutes and NGO's, such as hospitals, micro data about each and every person (e.g. medical history) for statistical or research purposes. However, sensitive personal information may be disclosed in this process, due to the existence in the data of quasi-identifying attributes, or simply quasi-identifiers (QID), such as age, zip code, ID, Birthdate etc. An attacker can join the QID with external information, such as voting registration lists, to identify individual person's records. Existing privacy-preserving techniques focus on anonymizing personal data, which have a fixed schema with a small number of dimensions. Through generalization, bucketization and suppression, existing methods prevent attackers from identifying individual records. For example, should businesses trust their employees with the critical role of protecting sensitive corporate information? Industry analysts would probably say "never" and with good reason. According to one recent Forrester study, 80 percent of data security breaches involve insiders, employees or those with internal access to an organization, putting information at risk. The big challenge for companies today – particularly as email and the Internet make sharing and distributing corporate information easier than ever .For example, database users traditionally are assigned a database administrator (DBA) role or granted multiple system privileges. As companies continue to consolidate databases and streamline operations to maximize efficiency and the protection of data from external threats, this user- and role-based security model no longer complies with "need-to-know" security best-practices. Today, to help ensure the safety, integrity and privacy of corporate information, more companies are pursuing a comprehensive, multi-factored security approach. For example, in a database which is having large datasets with a high dimension data such as Customer personal data such as Customer ID, Address, Phone No., Account details, Purchase details etc., such database table should be secure. Such data when shown to outer .

LITERATURE SURVEY

PRIVACY-PRESERVING CONTEXT TO DATA MINING

Generally when people talk of privacy, they say .keep information about me from being available to others. It is this intrusion, or use of personal data in a way that negatively impacts someone's life, that causes concern. As long as data is not misused, most people do not feel their privacy has been violated. The problem is that once information is released, it may be impossible to prevent misuse. Utilizing this distinction, ensuring that a data mining would not enable misuse of personal information opens opportunities that complete privacy would prevent. To do this, technical and social solutions that ensure data will not be released. The same basic concerns also apply to collections of data.

PRIVACY-PRESERVING CONTEXT TO DATA PUBLISHING

These techniques tend to study different transformation methods associated with privacy. These techniques include methods such as suppression, randomization, k-anonymity, and l-diversity. Another related issue is how the perturbed data can be used in conjunction with by means of classical data mining methods such as association rule mining. Other related problems include that of determining privacy-preservation methods to keep the underlying data useful (utility-based methods), or the problem of studying the different definitions of privacy, and how they compare in terms of effectiveness in different scenarios. In the most basic form of PPDP, the data holder has a table of the form D(Explicit Identifier, Quasi Identifier, Sensitive Attributes, Non-Sensitive Attributes), where Explicit Identifier is a set of attributes, such as name and social security number (SSN), containing information that explicitly identifies record owners; Quasi Identifier is a set of attributes that could potentially identify record owners; most sensitive Attributes consist of sensitive person-specific information such as disease, salary, and disability status; and Non-Sensitive Attributes contains all attributes that do not fall into the previous three categories. Most works assume that the four sets of attributes are disjoint.

ANONYMIZATION TECHNIQUES

I. GENERALIZATION

Generalization replaces quasi-identifier values with values that are less-specific but semantically consistent. Then, all quasi-identifier values in a group would be generalized to the entire group extent in the QID space. In order for generalization to be effective, records in the same bucket must be close to each other so that generalizing the records would not lose too much information, in high-dimensional data, most data points have similar distances with each other possible. This is main problem of generalization that prevents effective analysis of attribute correlations.

DISADVANTAGES OF GENERALIZATION

The main two disadvantages with generalization are:

1) It fails on high-dimensional data due to the curse of dimensionality.

2) It loses considerable amount of information, especially for high dimensional Data due to the uniform-distribution assumption.

II. BUCKETIZATION

Bucketization is to partition the tuple in T into buckets, and then to separate the sensitive attribute from the non-sensitive ones by randomly permuting the sensitive attribute values within each bucket. The sanitized data then consists of the buckets with permuted sensitive values. Bucketization first partitions tuples in the table into buckets and then separates the quasi identifiers with the sensitive attribute by randomly permuting the sensitive attribute values in each bucket. However, their approach assumes a clear separation between QIs and SAs. In addition, because the exact values of all QIs are released, membership information is disclosed.

DISADVANTAGES OF BUCKETIZATION:

It has better data utility than generalization, it has several disadvantages.

1) Bucketization does not prevent membership disclosure.

2) Bucketization requires a clear separation between QIs and SAs.

SYSTEM ARCHITECTURE

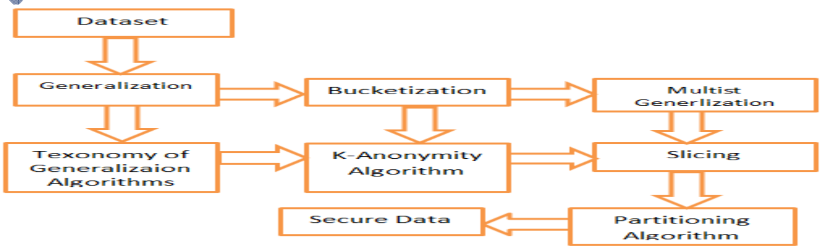

Figure 1: System Architecture

Algorithmic Procedure

1] Extract the data set from the database.
2] Anonymity process divides the records into two.
3] Interchange the sensitive values.
4] Multistep values generated and displayed.
5] Attributes are combined and secure data Displayed.

.

SLICING ALGORITHM

Slicing algorithm consists of three phases: attribute partitioning, column generalization, and tuple partitioning.

ATTRIBUTE PARTITIONING

This algorithm partitions attributes so that highly correlated attributes are in the same column. This is good for both utility and privacy. In terms of data utility, grouping highly correlated attributes preserves the correlations among those attributes. In terms of privacy, the association of uncorrelated attributes presents higher identification risks than the association of highly correlated attributes because the associations of uncorrelated attribute values is much less frequent and thus more identifiable.

TUPLE PARTITIONING

The algorithm maintains two data structures, the algorithm takes one scan of each tuple t in the table t to find out all tuples that match b and record their matching probability p(t, B) and the distribution of candidate sensitive values d(t, B) which are added to the list l(t). We have obtained, for each tuple t, the list of statistics L (t) about its matching buckets. A final scan of the tuples in t will compute the p (t, b) values based on the law of total probability.

COLUMN GENERALIZATION:

First column generalization may be required for identity/membership disclosure protection. If a column value is unique in a column, a tuple with this unique column value can only have one matching bucket. This is not good for privacy protection, as in the case of generalization bucketization where each tuple can belong to only one equivalence class/bucket

CONCLUSION

Slicing overcomes the limitations of generalization and bucketization and preserves better utility while protecting against privacy threats. We illustrate how to use slicing algorithms like attribute and tuple partitioning, column generalization to prevent attribute disclosure and membership disclosure. Our experiments show that slicing preserves better data utility than generalization and is more effective than bucketization.

REFERENCES

[1] D. Kifer and J. Gehrke, "Injecting Utility into Anonymized Data Sets," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 217-228, 2006.

[2] D.J. Martin, et.al, "Worst-Case Background Knowledge for Privacy- Preserving Data Publishing," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 126-135, 2007.

[3] Gabriel Ghinita, Member IEEE, Panos Kalnis, Yufei Tao, "Anonymous Publication of Sensitive Transactional Data" in Proc. Of IEEE Transactions on Knowledge and Data Engineering February 2011 (vol. 23 no. 2) pp. 161-174.

[4] G.Ghinita, Y. Tao, and P. Kalnis, "On the Anonymization of Sparse High-Dimensional Data," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 715-724, 2008.

[5] J. Xu, et. al,"Utility- Based Anonymization Using Local Recoding," Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 785-790, 2006.

[6] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan,"Incognito: Efficient Full-domain k-Anonymity," in Proc. of ACM SIGMOD, 2005, pp. 49– 60.

[7] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan,"Mondrian Multidimensional k-Anonymity," in Proc. of ICDE, 2006.

[8] R. J. Bayardo and R. Agrawal, "Data Privacy through Optimal k- Anonymization," in Proc. of ICDE, 2005, pp. 217–228.

[9] T. Li and N. Li, "Injector: Mining Background Knowledge for Data Anonymization," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 446-455, 2008.
[10] T. Li and N. Li, "On the Trade-off between Privacy and Utility in Data Publishing," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 517-526, 2009.
{11] X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 139-150, 2006.
[12] Y. He and J. Naughton, "Anonymization of Set-Valued Data via Top-Down, Local Generalization," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 934-945, 2009.
[13] Y. Xu, K. Wang, et.al,"Anonymizing Transaction Databases for Publication" Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 767-775, 2008.