ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

ANONYMIZING TRANSACTION DATA PUBLICATION USING SLICING

Dr.K.P.Kaliyamurthie, D.Parameswari

Professor & Head, Dept of IT, Bharath University, Chennai-600073, India

Asso. Professor, Dept. of MCA, Jerusalem college of Engineering, Chennai-600100, India

ABSTRACT : Slicing technique has been proposed as a mechanism for protecting privacy in microdata publishing. To hide certain customer specific information while releasing the transaction data to a third party. Several anonymization techniques have been used for data publishing. Generalization technique loses considerable amount of information, especially for high dimensional data. Bucketization technique does not have a clear separation between quasi-identifying attributes and sensitive attributes. Slicing preserves better data utility than generalization and bucketization. This technique partitions the data both horizontally and vertically. Slicing can handle high dimensional data.

KEYWORDS*:* Slicing, Generalization, Bucketization

I. INTRODUCTION

Data mining is the process of extracting useful, interesting and previously unknown information from large data sets. The success of data mining relies on the availability of high quality data and effective information sharing. The collection of digital information by governments, corporations and individuals has created an environment that facilitates large scale data mining and data analysis. Moreover, driven by mutual benefits, or by regulations that require certain data to be published, there is a demand for sharing data among various parties. There are several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns.

The detailed data in its original form often contain sensitive information about individuals and sharing such data could potentially violate individual privacy. The general public expresses serious concerns on their privacy and the consequences of sharing their person-specific information. The current privacy protection practice primarily relies on policies and guidelines to restrict the types of publishable data, and agreements on the use and storage of sensitive data. The data publisher collects data from record owners and releases the collected data to a data miner or the public, called the data recipient, who will then conduct data mining on the published data.

The data holders who collect data from record owners may be of two types of models such as untrusted model and trusted model. In the untrusted model, the data holder is not trusted and may attempt to identify sensitive information from record owners. Various cryptographic solutions, anonymous communications and statistical methods were proposed to collect records anonymously from their owners without revealing the owners identity. In the trusted model, the data holder is trustworthy and record owners are willing to provide their personal information to the data holder, however the trust is not transitive to the data recipient.

Anonymization seeks to hide the identity and/or the sensitive data of record owners, assuming that sensitive data must be retained for data analysis. Various anonymization techniques are generalization, bucketization, anatomization, permutation and perturbation. Generalization replaces some values with a parent value in the taxonomy of an attribute. Each generalization operation hides some details in QID. For a categorical attribute, a specific value can be replaced with a general value according to a given taxonomy. Bucketization is used to replace the values within the sensitive attributes. In the SA bucket, values in each column are randomly permutated to break the linking between different columns. Anatomization and permutation de-associate the correlation between QID and sensitive attributes by grouping and shuffling sensitive values in a qid group. Perturbation distorts the data by adding noise, aggregating values, swapping values, or generating synthetic data based on some statistical properties of the original data.

Slicing partitions the data set both vertically and horizontally. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes. Each column contains a subset of attributes that are highly correlated. Horizontal partitioning is done by grouping tuples into buckets. Finally, within each bucket, values in each column are randomly permutated (or sorted) to break the linking between different columns.Slicing preserves better data utility than generalization and can be used for membership disclosure protection and it can handle high-dimensional data. Also slicing preserves better utility than generalization and is more effective than bucketization involving the sensitive attribute.

II.PREVIOUS RESEARCH

Slicing partitions the data set both vertically and horizontally. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes. Each column contains a subset of attributes that are highly correlated. Horizontal partitioning is done by grouping tuples into buckets. Finally, within each bucket, values in each column are randomly permutated (or sorted) to break the linking between different columns. Slicing preserves better data utility than generalization and can be used for membership disclosure protection and it can handle high-dimensional data. Also slicing preserves better utility than generalization and is more effective than bucketization involving the sensitive attribute.[1]

Access to the published data should not enable the adversary to learn anything extra about any target victim compared to no access to the database even with the presence of any adversary's background knowledge obtained from other sources. Privacy preserving data publishing is a study of eliminating privacy threats while at the same time, preserving useful information in the released data for data mining. A task of the utmost importance is to develop methods and tools for publishing data in a hostile environment so that the published data remain practically useful while individual privacy is preserved. This undertaking is called privacy-preserving data publishing (PPDP).[2]

The data publisher collects data from record owners and releases the collected data to a data miner or the public, called the data recipient, who will then conduct data mining on the published data. The detailed data in its original form often contain sensitive information about individuals and sharing such data could potentially violate individual privacy. The general public expresses serious concerns on their privacy and the consequences of sharing their person-specific information. The current privacy protection practice primarily relies on policies and guidelines to restrict the types of publishable data, and agreements on the use and storage of sensitive data. Anonymization seeks to hide the identity and/or the sensitive data of record owners, assuming that sensitive data must be retained for data analysis.[3]

In real-life application, such absolute privacy protection is impossible due to the presence of the adversary's background knowledge. The attack occurs are record linkage, attribute linkage and table linkage respectively. The first category considers that a privacy threat occurs when an adversary is able to link a record owner to a record in a published data table. K-anonymity is used to prevent record linkage attack. If one record in the table has some value qid, at least k-1 other records also have the value qid. In other words, the minimum equivalence group size on QID is at least k. A table satisfying this requirement is called k-anonymous. In a k-anonymous table, each record is indistinguishable from at least k-1 other records with respect to QID.[4]

In real-life application, such absolute privacy protection is impossible due to the presence of the adversary's background knowledge. The attack occurs are record linkage, attribute linkage and table linkage respectively. The first category considers that a privacy threat occurs when an adversary is able to link a record owner to a sensitive attribute in a published data table. ℓ-diversity is used to prevent attribute linkage. The ℓ-diversity requires every qid group to contain at least ℓ "well-represented" sensitive values.[5]

Each transaction is an arbitrary set of items chosen from a large universe. Detailed transaction data provides an electronic image of one's life. This has two implications, one transaction data are excellent candidates for data mining research. Two, use of transaction data would raise serious concerns over individual privacy. Therefore before transaction data is released for data mining, it must be made anonymous so that data subjects cannot be re-identified.[6]

III. HYPOTHESES

System design is the process or art of defining the architecture, components, modules, interfaces and data for a system to satisfy specified requirements. One could see it as the application of systems theory to product development.

H1: MEASURES OF CORRELATION

Mean square contingency coefficient is used for measuring correlation because most of our attributes are categorical. The mean-square contingency coefficient between A_1 and A_2 is defined as

$$\emptyset^2(A_1, A_2) = \frac{1}{\min\{d_1, d_2\} - 1} \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j}$$

It can be shown that $0 \leq \emptyset^2(A_1, A_2) \leq 1$.

H2: **ATTRIBUTE CLUSTERING**

After computing the correlations for each pair of attributes we use clustering to partition attributes into columns. Each attribute is a point in the clustering space. The distance between two attributes in the clustering space is defined as $d(A_1, A_2) = 1 - \emptyset^2(A_1, A_2)$, which is in between of 0 and 1. Two attributes that are strongly correlated will have a smaller distance between the corresponding data points in our clustering space.

IV. RESEARCH METHOD

In the existing system several anonymization techniques such as generalization and bucketization, have been designed for privacy preserving microdata publishing. While examining generalization loses considerable amount of information, especially for high-dimensional data. Bucketization does not prevent membership disclosure and does not apply for data that have a clear separation between quasi-identifying attributes and sensitive attributes. To avoid these slicing a novel technique is used. Slicing partitions the data both horizontally and vertically. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes.

4.1. Hypotheses Testing

4.1.1. MEASURES OF CORRELATION (H1)

Mean square contingency coefficient is used for measuring correlation because most of our attributes are categorical. Given two attributes A_1 and A_2 with domains { v_{11} , v_{12} ,…., v_{1d_1} } and {v_{21},v_{22},…..,v_{2d_2}}, respectively. Their domain sizes are thus d_1 and d_2, respectively. The mean-square contingency coefficient between $A_1$and A_2 is defined as

$$\emptyset^2(A_1, A_2) = \frac{1}{\min\{d_1, d_2\} - 1} \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \frac{(f_{ij} - f_i f_j)^2}{f_i f_j}$$

Here, f_i and f_j are the fraction of occurrences of v_{1i} and v_{2j} in the data, respectively. $f_{ij}$is the fraction of occurrences of v_{1i} and v_{2j}in the data. Therefore, f_i and f_j are the marginal totals off_{ij}: $f_i = \sum_{j=1}^{d_2} f_{ij}$ and $f_j = \sum_{i=1}^{d_1} f_{ij}$. It can be shown that $0 \leq \emptyset^2(A_1, A_2) \leq 1$.

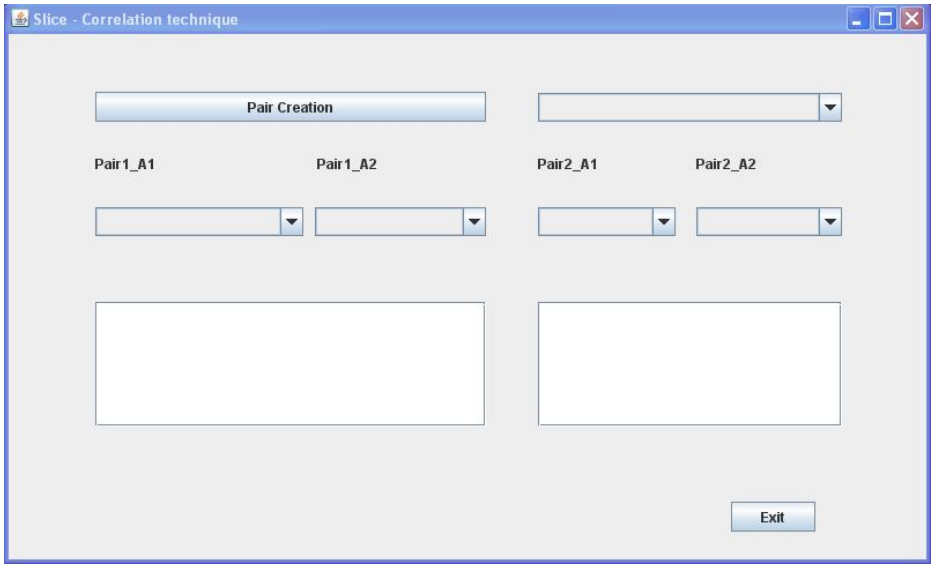
4.1.2 **ATTRIBUTE CLUSTERING** (H2)

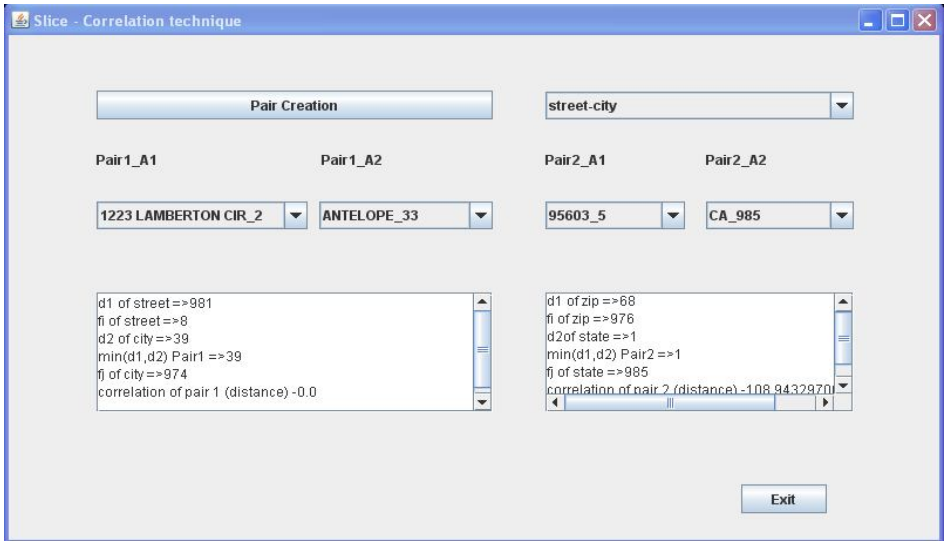
After computing the correlations for each pair of attributes we use clustering to partition attributes into columns. Each attribute is a point in the clustering space. The distance between two attributes in the clustering space is defined as $d(A_1, A_2) = 1 - \emptyset^2(A_1, A_2)$, which is in between of 0 and 1. Two attributes that are strongly correlated will have a smaller distance between the corresponding data points in our clustering space.

V.THE RESULTS

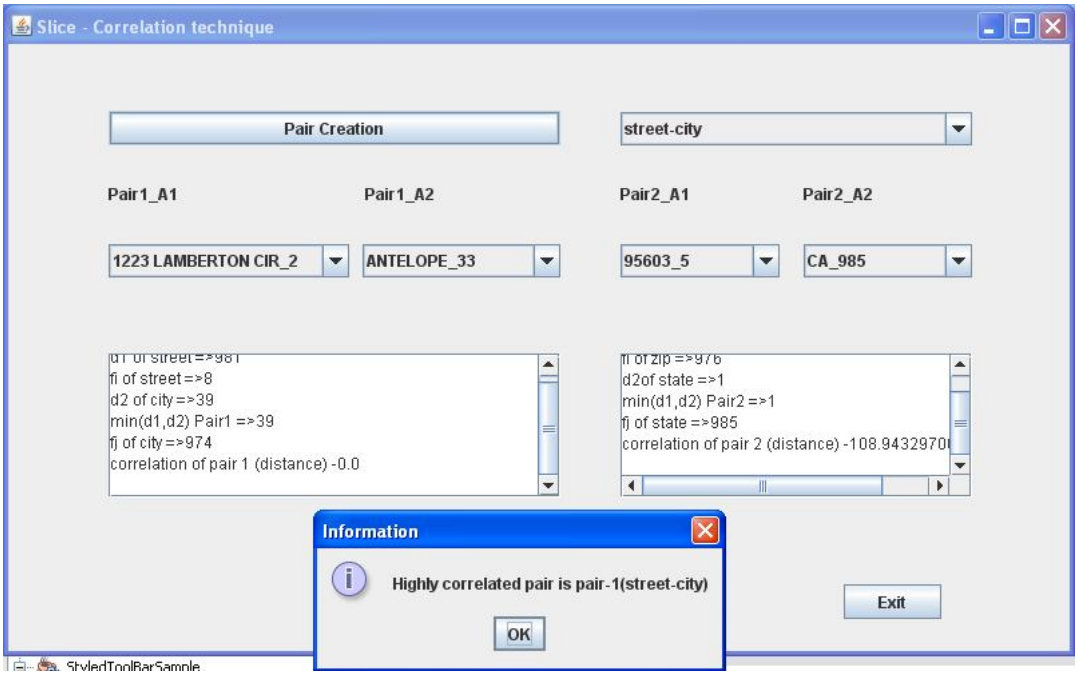
In this section describes the results of measures of correlation and attribute clustering.

5.1. Results of Measures of Correlation H1





5.2. The Results of Attribute Clustering H2



VI. SUMMARY AND CONCLUDING REMARKS

In this paper when there exist more than one categorical attribute that can be chosen to perform partitioning; when there is a need to cluster multiple values of the attribute before partitioning on it. This algorithm has been tested on some small datasets. It is proved that this algorithm can select an appropriate partitioning attribute as well as form appropriate partitioning on multiple valued attributes.

REFERENCES

[1] Tiancheng Li, Ninghui Li, Senior Member, IEEE, Jian Zhang, Member, IEEE, and Ian Molloy, "Slicing: A New Approach For Privacy Preserving Data Publishing," *IEEE Trans. Knowledge and Data Eng.*, vol. 24, no. 3, March 2012.

[2] C. C. Aggarwal, J. Pei, and B. Zhang. On privacy preservation against adversarial data mining. In *Proc. of the 12th ACM International Conference on Knowledge Discovery and Data Mining* (SIGKDD), Philadelphia, PA, August 2006

[3]Z. Yang, S. Zhong, and R. N.Wright. Anonymity-preserving data collection. In Proc. of the 11th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 334–343, 2005.

[4] L. Sweeney. k-Anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

[5] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. In *Proc. of the 22nd IEEE International Conference on Data Engineering (ICDE)*, Atlanta, GA, 2006.

[6]Y. Xu, K. Wang, A.W.-C. Fu, and P.S. Yu, "Anonymizing Transaction Databases for Publication," *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 767-775, 2008.