International Journal of
Engineering Research and Science & Technology

IJERST

www.ijerst.com

Research Paper

MICRO DATA PUBLISHING BY SLICING APPROACH WITH PRIVACY PRESERVATION

N V Kalyani[1]* and K Sujatha[2]

*Corresponding Author: **N V Kalyani** ✉ kalyanil9cs1129@gmail.com*

With the advent of new trends in the present environment the anomynization techniques are not dealing with privacy preservation and multidimensional data sets in a perfect manner. For that reason some valuable information may be lost. In this paper, we propose slicing behavior in the fashion of overlapping with the duplication of sensitive data for hiding the data. These slicing techniques are given the effective utility of data and protect the data. In the slicing columns grouping also place major role for publishing the micro data with the columns linking behavior.

Keywords: Micro data publishing, Data anonymization, Privacy preservation, Slicing, Sensitive Data

INTRODUCTION

In this fast growing world there is a need for data mining tools to analyze the massive amounts of data bases. The database includes the particulars regarding either health records, personal details or other important data related to individuals (Aggarwal, 2005). The data mining tools include mathematical models, statistical models and machine learning models (neural networks and decision trees). Coming to the organizations, they want to protect their own publishing transactional dataset from unwanted threats and unauthorized access. Here the publishing of data is usually in the form of micro data formats. Which contain individuals identity.

The synonym for privacy is anonymity which means to hidden completely the individual's identity. There are several anonymity variations for one's identity. Most well known variation in the anonymity is generalization (Machanavajjhala, 2006). In this generalization several methods can be used like k-anonymity, adaptative clustering, etc. Another technique used for micro data publishing is bucketization. In the annonimization techniques table attributes are partitioned into three different ways. The first way is identifiers like social security numbers, names, etc. In the second way, There are Quasi Identifiers (QI) known as attributes like zip code, birth date, sex, etc. In the last way, there is sensitive data (SA) and unknown attributes like salary and disease, etc. In privacy preservation (Sweeney, 2002; Tiangcheng, 2012) above two techniques are used in common but they show some of the drawbacks while dealing with multi dimensional

[1]   M. Tech Student, Dept. of CSE, Malineni Lakshmaiah Engineering College, Singarayakonda, Prakasam(dt.), AP.
[2]   Assoc. Prof., Department of CSE, Malineni Lakshmaiah Engineering College, Singarayakonda, Prakasam(dt.), AP.

data and lose some useful information. Generalization and bucketization do not display the correct data regarding micro data publishing.

SLICING BEHAVIOR

For dealing with high dimensional data and privacy preservation (Samarati, 2001) purposes in micro data publishing alternative technique in annonimyzation (Inan, 2009) is slicing. In slicing technique, there is a partition the given data set both in horizontally and vertical manners. In vertical partitions grouping the attributes into columns contains the high correlation among them (Vijay kumar, 2013). Horizontal partitions based on combining the given data set tuples into buckets. In this partition each column randomly

Table 1: An original Micro Data Table and Various Slicing Behaviors			
Age	**Sex**	**Zip code**	**Disease**
(a) The original table			
20	M	28603	Dengue
20	F	28603	Feaver
23	F	28602	Feaver
41	F	28602	cold
43	M	28402	Feaver
50	M	28402	Dengue
50	M	28404	Dengue
54	F	28404	Asthma
(b) One –attribute-per-column slicing			
20	M	28603	Dengue
20	F	28603	Feaver
23	F	28602	Feaver
41	F	28602	Cold
43	M	28402	Feaver
50	M	28402	Dengue
50	M	28404	Dengue
54	F	28404	Asthma
(Age, Sex)		**(Zip code, Disease)**	
(c) Sliced table			
(20, M)		(28603, Dengue)	
(20, F)		(28603, Feaver)	
(23, F)		(28602, Feaver)	

Table 1 (Cont.)	
(Age, Sex)	**(Zip code, Disease)**
(41, F)	(28602, cold)
(43, M)	(28402, Feaver)
(50, M)	(28402, Dengue)
(50, M)	(28404, Dengue
(54, F)	(28404, Asthma)
(d) Overlapping sliced table	
(20, M, Dengue)	(28603, Dengue)
(20, F, Feaver)	(28603, Feaver)
(23, F, Feaver)	(28602, Feaver)
(41, F, cold)	(28602, cold)
(43, M, Feaver)	(28402, Feaver)
(50, M, Dengue)	(28402, Dengue)
(50, M, Dengue)	(28404, Dengue)
(54, F, Asthma)	(28404, Asthma)

sorts to break the linking between the different columns (Li and Li, 2009).

Here in Table 1(a) shows an example table for micro data publishing and shows the behavior of table after applying various anonymization (slicing) techniques (Koudas, 2007; Xiao, 2006). From this table, attributes are identified as QI (Age, Sex, Zip code) and SA (Disease) attributes. Table 1(b) shows the behavior of one-attribute-per-column behavior. In this, individual column behaves as a table. Table 1(c) shows the behavior of slicing technique partitioned into horizontal and vertical manner. The table contains two columns: the first column contains {Age, Sex} and the second column contains {Zip code, Disease}. The given table is partitioned into horizontal manner like buckets. Each bucket has four tuples. In Table 1(c) each bucket in two columns are randomly permuted means bucket columns linking is hidden. In this slicing better privacy preserving can be possible because it breaks the uncorrelated attribute associations. With this slicing technique get the better utility of data and reduces the dimensionality of data rather than previous two techniques.

For maintaining good privacy preservation along with reducing the dimensionality, duplication of the SA attributes is needed. The new mechanism called overlapping slicing has the duplication of sensitive data. As shown in Table 1(d) overlapping slicing partitioned in the same manner of slicing technique Table 1(c), one simple difference is sensitive attribute disease can be duplicated in both columns. Here two columns are in the manner of (18, M, Dengue), (18, F, Feaver), (23, F, Feaver), (41, F, cold). These are randomly permuted and linking with the (28603, Dengue), (28603, Feaver), (28602, Feaver),

(28602, cold). This slicing variation deals with the data utility and multi dimensional data in better manner rather than the other slicing behaviors. With this duplication of data hiding of sensitive data in effective manner for the purpose of micro data publishing. This overlapping slicing technique reduces the tradeoff between the utility and privacy of the data.

CONCLUSION

The slicing anonymization technique for hiding the data between the columns is the main criteria. By comparing the generalization and bucketization, the proposed slicing and overlapping slicing gives more data privacy and reduces the dimensionality of data sets for the purpose of micro data publishing purpose. The randomly permuted mechanism is used between the columns linking. This slicing behavior can be efficient in hiding one's identity.

FUTURE WORK

By applying the overlapping of slicing sensitive data may be duplicated for the purpose of hiding. In future our intention is to reduce the multidimensional data sets up to single dimension data set that may reduce the threats into micro data publishing. In the present days so many anonymization techniques are involved for the purpose of micro data publishing with privacy preservation. Up to now we are using the grouping technique in random permutation. In some of the situations by using this random permutation, we are not getting the perfect results. For that reason applying the continuous random mechanism in the anonimyzation techniques is to get the effective micro data publishing with privacy preservation.

REFERENCES

1. Aggarwal C (2005), "On k -Anonymity and the Curse of Dimensionality," *Proc. Int'l Conf. Very Large Data Bases (VLDB),* pp. 901-909.

2. Inan A, Kantarcioglu M, and Bertino E (2009), "Using Anonymized Data for Classification," *Proc. IEEE 25th Int'l Conf. Data Eng. (ICDE),* pp. 429-440.

3. Vijay Kumar P and Manika chzian R (2013), "Privacy Preservation Using Randomized Attribute Selection Based On Knowledge Hiding", *The International Journal Of Engineering And Science (IJES),* Vol. 2, Issue 4, pp. 53-57.

4. Koudas N, Srivastava D, Yu T, and Zhang Q (2007), "Aggregate Query Answering on Anonymized Tables," *Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE),* pp. 116-125.

5. Li T and Li N (2009), "On the Tradeoff between Privacy and Utility in Data Publishing," *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD),* pp. 517-526, 2009.

6. Machanavajjhala A, Gehrke J, Kifer D, and Venkitasubramaniam M (2006), "l-diversity: Privacy Beyond K-Anonymity," *ProcInt'lConf. Data Eng (ICDE),* p. 24.

7. Samarati P (2001), "Protecting Respondent's Privacy in Microdata Release," *IEEE Trans. Knowledge and Data Eng.,* Vol. 13, No. 6, pp. 1010-1027.

8. Sweeney L (2002), "K-Anonymity: A Model for Protecting Privacy," *Int'l J. Uncertainty Fuzziness and Knowledge-Based Systems,* Vol. 10, No. 5, pp. 557-570.

9. Tiancheng Li, Ninghui Li, Jian Zhang, and Ian Molloy (2012), "Slicing: A New Approach for Privacy Preserving Data Publishing", *IEEE Transactions On Knowledge And Data Engineering*, Vol. 24, No. 3.

10. Wong R C W, Li J, Fu A W C and Wang K (2006), "K-Anonymity: An Enhanced K-Anonymity Model for Privacy Preserving Data Publishing," *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD),* pp. 754-759.

11. Xiao X and Tao Y (2006), "Anatomy: Simple and Effective Privacy Preservation," *Proc. Int'l Conf. Very Large Data Bases (VLDB),* pp. 139-150.

12. Xiao X and Tao Y (2006), "Anatomy: Simple and Effective Privacy Preservation," *Proc. Int'l Conf. Very Large Data Bases (VLDB),* pp. 139-150.

International Journal of Engineering Research and Science & Technology

Hyderabad, INDIA. Ph: +91-09441351700, 09059645577

E-mail: editorijlerst@gmail.com or editor@ijerst.com

Website: www.ijerst.com