Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 4, April 2014, pg.1393 – 1400

RESEARCH ARTICLE

EFFICIENT TECHNIQUES FOR PRESERVING MICRODATA USING SLICING

Ms. Neenu Varghese[1], Mr. Avanish Kumar Singh[2]

[1]Department of Computer Science and Engineering, Calicut University, India
[2]Department of Computer Science and Engineering, Calicut University, India

[1] neenumariya10@gmail.com
[2] avnpt@rediffmail.com

***Abstract*—Privacy preserving publishing is the kind of techniques to apply privacy to collected vast amount of data. One of the recent problem prevailing is in the field of data publication. The data often consist of personally identifiable information so releasing such data consists of privacy problem. Several anonymization techniques such as generalization and bucketization have been designed for privacy preserving and microdata publishing. But the problem is that generalization loses considerable amount of information and bucketization does not prevent membership disclosure. So we proposed a novel technique called slicing, which partitions the data both vertically and horizontally. An efficient algorithm is also developed for performing slicing that obeys l-diversity requirement. An advantage of slicing is that it can handle high dimensional data. Our experiment confirms that slicing preserves better utility than generalization and is more effective than bucketization in workloads involving the sensitive attributes.**

Keywords*— *Privacy preserving; l-diversity; Data Publishing; Microdata

I. INTRODUCTION

Privacy preservation in microdata publishing is a series issue in recent years. Today most of the fields in industry need to publish microdata. The microdata consist of information about an individual or anything. Microdata contain records each of which contains information about an individual entity such as a person, a household or an organization. The main objective is that individual's information is not revealed as well as the data must be useful. Many microdata anonymization techniques have been proposed and the most popular ones are generalization and bucketization. In both approaches attributes are partitioned into 3 categories[1].

- Some attributes are identifiers that can uniquely identify an individual such as name or social security number.

- Some attributes are quasi identifiers whose values when taken together can potentially identify an individual. Example includes zip code, birth date, and gender.

- Some attributes are sensitive attributes this kind of attributes are unknown to the opponent and are considered sensitive such as disease and salary

Privacy preserving data analysis and data publishing has received considerable information promising that it can share the data providing that individual privacy is preserved.

II. RELATED WORKS

In this chapter we discuss about the related work done in privacy preserving microdata publishing. Several anonymization techniques such as generalization and bucketization have been designed for privacy preserving and microdata publishing. But the problem is that generalization loses considerable amount of information and bucketization does not prevent membership disclosure. Two main problems of generalization are:

1. Fails on high-dimensional data due to the curse of dimensionality.

2. Too much information loss due to uniform distribution.

Limitations of Bucketization

1. Does not prevent membership disclosure.

2. Requires a clear separation between QIs and SAs.

3. Breaks the attribute correlations between the QIs and the SAs by separating the SA from the QI attributes.

R chen et al proposed privacy preserving data publishing. In today's information age more and more information are evolving. There is information from governments, co, operations, healthcare records and individual's etc. There is demand for exchange and publication of data among varies parties. The task is to develop methods so that the published data remains useful and privacy is preserved. This undertaking is called privacy preserving data publishing(PPDP).It is a promising approach to information sharing, while preserving individual privacy and protecting sensitive information[1].

D Kifer proposed l diversity: privacy beyond k anonymity. Two main attacks is prevailing under the k anonymized dataset. Firstly the attacker can discover the value of sensitive attributes when there is diversity in those sensitive attributes; Secondly k anonymity does not guarantee privacy against attackers using background knowledge, so in order to avoid this l-diversity is proposed [4].

N li et al proposed t-closeness privacy beyond k anonymity and l –diversity. Recent work shows that k-anonymity cannot prevent attribute disclosure and also l-diversity has no of limitations. So proposed a novel method called T-closeness [3].

III. BASIC IDEA OF SLICING

In this paper we present a new data anonymization technique called slicing to improve the current state of art. Slicing is the method of partitioning the data horizontally and vertically. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes. Horizontal partitioning is done by

grouping tuples into buckets. The main idea of slicing is to preserve the association within each column. It reduces dimensionality of data and preserves better data utility than bucketization and generalization.

A. Overview

The overall method of slicing has been discussed above. The original microdata consist of quasi identifying values and sensitive attributes. In figure 1 patient data in a hospital. The data consists of Age, Sex, Zip code, disease. Here the QI values are {age, sex, zip code} and the sensitive attribute is {disease}.A generalized table replaces values [1].

Age	Sex	Zip code	Disease
22	M	47906	Cancer
22	F	47906	Thyroid
33	F	47906	Thyroid
52	F	47906	Diabetes
54	M	47906	Thyroid
60	M	47906	Cancer
60	F	47906	Cancer

Fig 1.orginal micro data

In generalization there are several recordings. The recoding that preserves the most information is "local recoding". In local recoding first tuples are grouped into buckets and then for each bucket, one replaces all values of one attribute with a generalized value, because same attribute value may be generalized differently when they appear in different buckets [2].

Age	Sex	Zip code	Disease
[20-52]	*	4790*	Cancer
[20-52]	*	4790*	Thyroid
[20-52]	*	4790*	Thyroid
[20-52]	*	4790*	Cancer
[54-64]	*	4790*	Cancer
[54-64]	*	4790*	Nausea
[54-64]	*	4790*	Cancer
[54-64]	*	4790*	Thyroid

Fig 2.Generalized data

In bucketization also attributes are partitioned into columns, one column contains QI values and the other column contains SA values. In bucketization, one separates the QI and SA values by randomly permuting the SA values in each bucket. In some cases we cannot determine the difference between them two. So it has one drawback for microdata publishing. It also does not prevent membership disclosure [2].

Age	Sex	Zip code	Disease
22	M	47906	Thyroid
22	F	47906	Cancer
33	F	47905	Diabetes
52	F	47905	Thyroid
54	M	47902	Nausea
60	M	47902	Thyroid
60	M	47902	Cancer
64	F	47902	Cancer

Fig 3.Bucketized data

Slicing does not require the separation of those two attributes. The basic idea of slicing is to break the association cross columns, but to preserve the association within each column. This reduces the dimensionality of data and preserves better utility. Slicing partitions the dataset both horizontally and vertically. Data slicing can also handle high-dimensional data. It provides attribute disclosure protection [2].

(Age,Sex)	(Zipcode, disease)
(22,M)	(47905,Thyroid)
(22,F)	(47906,Cancer)
(33,F)	47905,Diabetes)
(52,F)	(47906,Thyroid)
(54,M)	(47904,Nausea)
(60,M)	(47902,Thyroid)
(60,M)	(47902,Cancer)
(64,F)	(47902,Cancer)

Fig 4.Sliced data

Randomly select the splitting point, means the data that is to be anonymized. Make it m private if it is not. Similarly we can continue selecting another splitting point. Finally the privacy fitness score is determined.

(Age,Sex)	(Zipcode,disease)
(22,*)	(4790*,Thyroid)
(22,*)	(4790*,Cancer)
(33,*)	(4790*,Diabetes)
(52,*)	(4790*,Thyroid)
(54,*)	(4790*,Nausea)
(60,*)	(4790*,Thyroid)
(60,*)	(4790*,Cancer)
(64,*)	(4790*,Cancer)

Privacy fitness score 180

Fig 5.privacy fitness score

B. SLICING ALGORITHM

Step 1: In the initial stage we consider a queue of buckets Q and a set of sliced buckets SB. Initially Q contains only one bucket which includes all tuples and SB is empty. So Q= {T}; SB=Ø.

Step 2: In each Iteration the algorithm removes a bucket from Q and splits the bucket into two buckets. Q=Q-{B};for l-diversity check(T,QU{B1,B2}USB,l);The main part of tuple partitioning algorithm is to check whether a sliced table satisfies l- diversity.

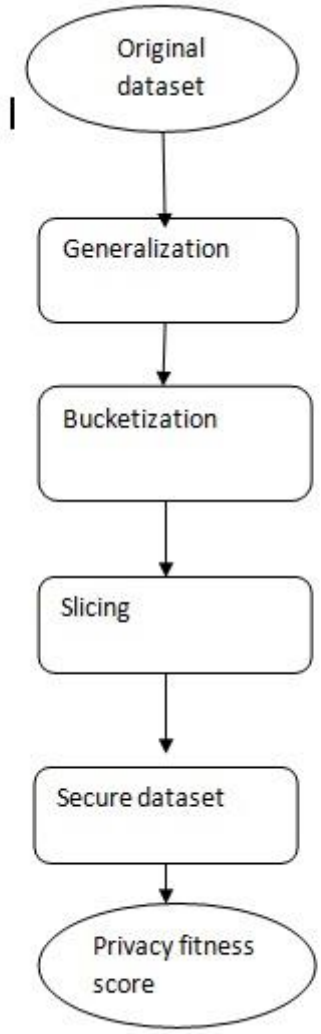
Step 3: In the diversity check algorithm for each tuple t, it maintains a list of statistics L[t] contains statistics about one matching bucket B. t€T, L[t] =Ø. The matching probability p (t, B) and the distribution of candidate sensitive values D (t, B).

Step 4: Q=QU {B1, B2} here two buckets are moved to the end of the Q.

Step 5: else SB=SBU {B} in this step we cannot split the bucket more so the bucket is sent to SB.

Step 6: Thus final results return SB, here when Q becomes empty we have computed the sliced table. The set of sliced buckets is SB .So, finally Return SB.

C. Architectural Diagram



IV. EXPERIMENTAL RESULTS

Here we see how doctors and patients registers using their details based on quasi-identifying attributes and sensitive attributes. After registering doctor logins and see the patient published details based on their diseases.

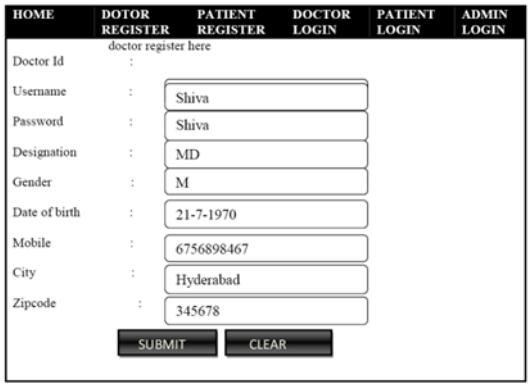


Fig 6.Doctor registration form

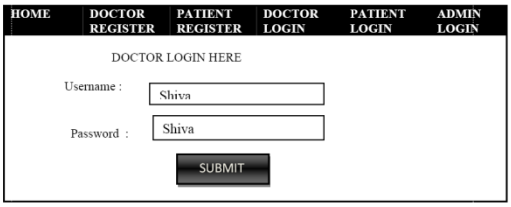


Fig 7.Doctor login form

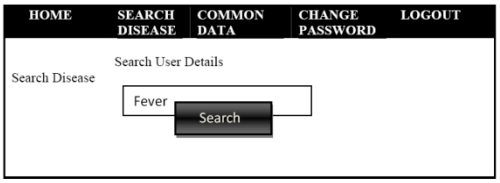


Fig 8.Data searching

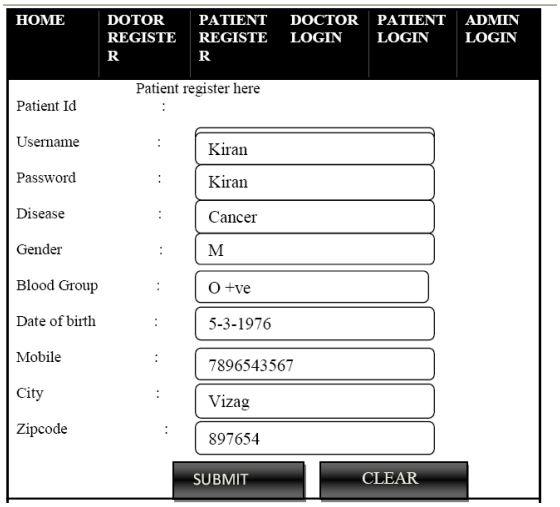


Fig 9.Patient Registration form

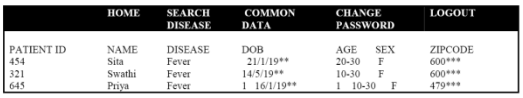


Fig 10.Published data

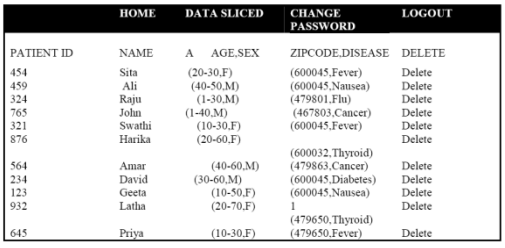


Fig 11.Sliced data

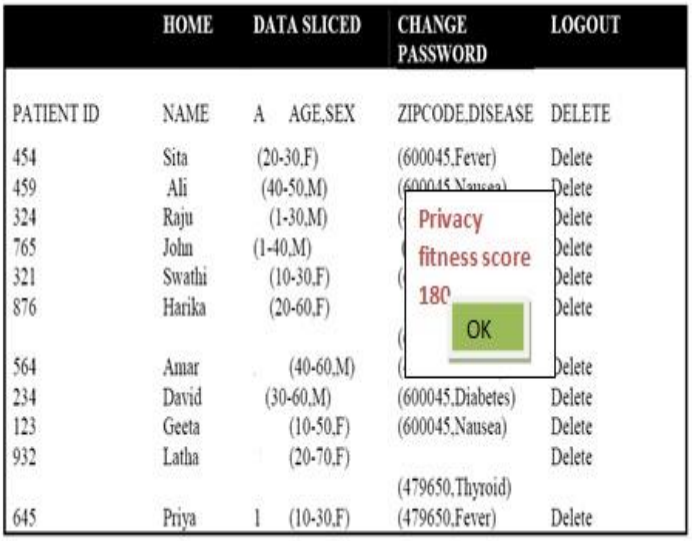


Fig. 12 privacy score

V. CONCLUSION

In this paper, we present a new anonymization method that is data slicing for privacy preserving and microdata publishing. Slicing overcomes the limitations of generalization and bucketization and preserves better utility while protecting against privacy threats .Data slicing is a promising technique for handling high dimensional data. By partitioning attributes into columns, privacy is protected.

REFERENCES

[1] Tiancheng Li, Ninghui Li, Senior Member, IEEE, Jia Zhang, Member, IEEE, and Ian Molloy "Slicing: A New Approach for Privacy Preserving Data Publishing" Proc. IEEE transactions on knowledge and data engineering, vol. 24, no. 3, march 2012.

[2] S. Goryczka, L. Xiong, and B. C. M. Fung, "m-privacy for collaborative data publishing," Emory University, Tech. Rep., 2011.

[3] N. Li and T. Li, "t-closeness: Privacy beyond k-anonymity and l-diversity,"in In Proc. of IEEE 23rd Intl. Conf. on Data Engineering(ICDE), 2007.

[4] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam,"l-diversity: Privacy beyond k-anonymity," in ICDE, 2006, p. 24.

