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 5, May 2014

A Narrative Approach for Data Preserving Techniques

K.S.Gangatharan*[1], M.S.Thanabal*[2]

[1]PG Scholar, Department of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India

[2]Associate Professor, Department of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul, Tamilnadu, India

ABSTRACT: In the recent year, the privacy takes major role to secure the data from various potential hackers. The privacy technique is used to avoid the stealing and reduce the leakage about the particular or individual information while the data are shared and realized to public. This paper focused for collaborative data publishing problem by anonymizing multiple data providers and generate the privacy to secure the data from new type of insider attacker. Varies approaches have been proposed to produce the privacy for anonymizing problem such as generalization, bucketization and slicing each of them has taken the solution of creating a privacy while data is publishing. Yet owing the possibilities of additional improvement, the system proposed in this paper takes the m-privacy and overlapping technique. This technique overcome the previous technique and shows the better result than the existing techniques.

KEYWORDS: Anonymization, Collaborative Publishing, Security, Privacy, Slicing.

I. INTRODUCTION

By using anonymization technique the data is modified and then released to the public. This process is known as the privacy preservation data publishing. The attributes are classified by three types which are Key attribute, quasi identifier and sensitive attribute. Key attribute which is represent a unique identification such as names, address, phone number and it always removed before publishing. Quasi-identifiers are segments of information that are not unique identifiers but well correlated with an entity they can be combined with other quasi-identifier to create a unique identifier. Example birth date, gender, which can be used link unionized dataset with other data. Last one is sensitive attributes example deceases, salaries, etc. from the fig.1 Consider the set of records t1, t2…. tn, which are provided by the provider. The record is a collection of some data. Before publishing the records to the public the anonymization technique is applied to the data, then it generate the subset of records t1, t2….. tn. Our goal is secure the original data or individual information from the different malicious user by using the anonymization when the data is published to the public. In the previous year varies techniques are used to private the data such as generalization and bucketization, slicing, m-privacy technique etc. But yet owing the additional improvement we are proposed the novel approach, which is the combination of generalization, bucketization, m-privacy and over lapping technique for private the data with high secure. It ensures the better privacy compared with the existing approaches.

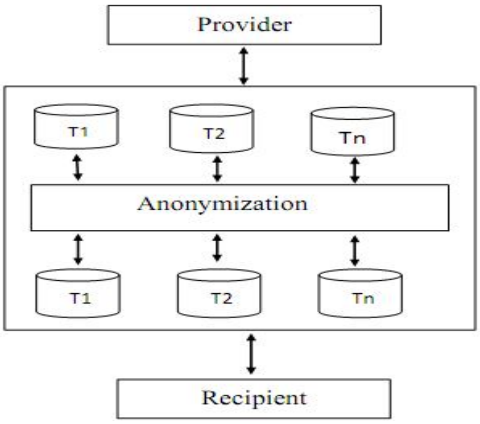
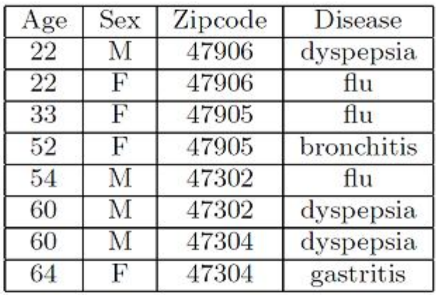


Figure. 1 Distributed data publishing with anonymization technique

II. Related work

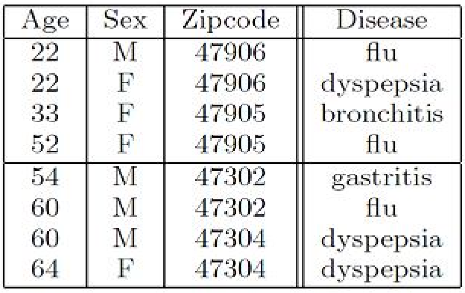
A. *Bucketization*

Bucketization is the process of the several records, grouping based on their sensitive values or non-sensitive attributes [1] [2]. The unequivocal sensitive values of the attributes are identified and sorted based on the frequencies in ascending order. After the sorting, the contiguous sensitive values are grouped into the congruent bucket. Only the buckets contain at ℓ distinct sensitive values which are kept after bucketing process completion. After the buckets are spliced into the group, the values of sensitive attributes are interrelated to its associated non-sensitive attributes or quasi identifier. The Table II illustrate the how to buckets are formed from Table I. The Table I consists some set of records R. Each record consists some set of attributes *d* with a set of values specified. Consider *d* = {a1, a2 ... an} be a set of attributes. Based on these set, identify the sensitive attributes and grouped into a set of buckets B = {B1, B2, B3….Bn}. Table II explains the sample dataset comprises with set of sensitive and non-sensitive attributes. In the dataset zip code, age, sex are non-sensitive attributes. Disease is a sensitive attribute. With the set of sensitive attributes obtained, the buckets are created in which it arbitrarily generates each set of sensitive attribute values among each set of bucket formed.



Age	Sex	Zipcode	Disease
22	M	47906	dyspepsia
22	F	47906	flu
33	F	47905	flu
52	F	47905	bronchitis
54	M	47302	flu
60	M	47302	dyspepsia
60	M	47304	dyspepsia
64	F	47304	gastritis

Table I. Original table



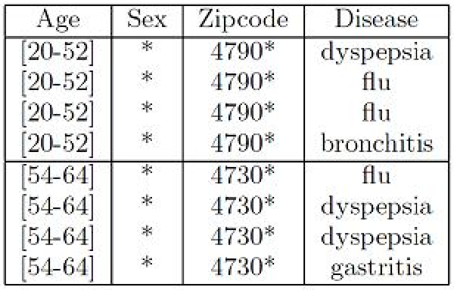
Age	Sex	Zipcode	Disease
22	M	47906	flu
22	F	47906	dyspepsia
33	F	47905	bronchitis
52	F	47905	flu
54	M	47302	gastritis
60	M	47302	flu
60	M	47304	dyspepsia
64	F	47304	dyspepsia

Table II. Bucketization table

In Table II , the sensitive attribute such as a disease has some values like the flu, dyspepsia, gastric, and bronchitis are interchanged its position and it is not related to its non-sensitive attribute such as age, sex and zip code. We can see the bucketization table differ from the original table. This is done, when the table or database are published to the public. The bucketization ensure the association of interrelated attributes are generates the privacy of the data while publishing to the public.

B. *Generalization*

Generalization is one of the general anonymized approaches [3]. It replace the QID values that are less specific, but values are consistent. In this approach at least two transactions in an individual group have a different values in separate column, then all the individual information about that item in the certain group is lost. While generalizing, the records would not lose too much information if the records in the same bucket must be close to each other. However, in high-dimensional data, most data values have similar distances with each other.



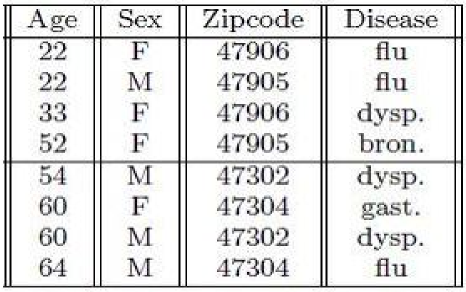
Age	Sex	Zipcode	Disease
[20-52]	*	4790*	dyspepsia
[20-52]	*	4790*	flu
[20-52]	*	4790*	flu
[20-52]	*	4790*	bronchitis
[54-64]	*	4730*	flu
[54-64]	*	4730*	dyspepsia
[54-64]	*	4730*	dyspepsia
[54-64]	*	4730*	gastritis

Table III. Generalization table

Table III describes the about the generalization approach. On that table there are two buckets which are spliced based on the sorting order of the age attribute. Then age attributes are generalized by the intervals, such the interval level is, the first value of the age level is starting value of the age attribute in the each bucket and the, last value of the age level is the last value of the age attribute in the each bucket, which mean 22 is the starting value of age attribute and the 52 is the ending value of the age attribute in the first bucket. which intervals is formed like as [20-52] then age attribute are consider as this interval values . The another quasi-identifier such as sex attributes values anonymized it means values are encrypted and another one quasi identifier such as values of the zip code are anonymized but the position of the values of the sensitive attribute values are not changed.

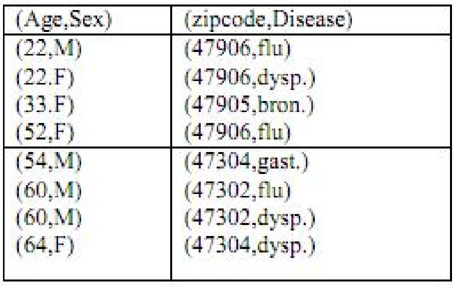
C. *Slicing*

Slicing first splits the attribute into columns and each column contains a subset of attributes. In the Table IV shows the one attribute per column slicing. The age attribute and zip code attribute are attribute of their own columns and the sex attribute is the subset of age attribute (or the sex column is the subset of age column) same as zip code attribute is the subset of the Disease attribute and the sex attribute is the subset of the zip code attribute and the zip code attribute is the subset of the sex attribute. Here encryption or anonymized technique is not used but tuples are grouped into a bucket. And the value of the sensitive attribute is not changed its position.



Age	Sex	Zipcode	Disease
22	F	47906	flu
22	M	47905	flu
33	F	47906	dysp.
52	F	47905	bron.
54	M	47302	dysp.
60	F	47304	gast.
60	M	47302	dysp.
64	M	47304	flu

Table IV. One- attribute per column slicing



(Age,Sex)	(zipcode,Disease)
(22,M)	(47906,flu)
(22.F)	(47906,dysp.)
(33.F)	(47905,bron.)
(52,F)	(47906,flu)
(54,M)	(47304,gast.)
(60,M)	(47302,flu)
(60,M)	(47302,dysp.)
(64,F)	(47304,dysp.)

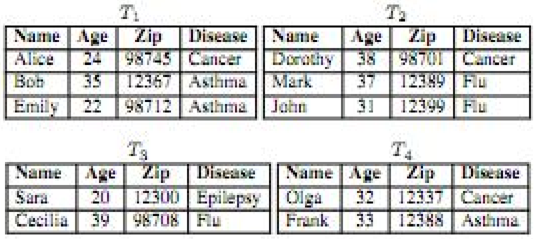
Table V. Sliced table

In the Table V describes about the slicing. Here the tuple of each bucket contains a value of age and value of the sex then it's it forms a one column and the values of subsets are changing its position such means interrelated to its association value attribute such as in table age and attribute values in the first bucket are (22,22,33,52) and (F,M,F,F) then it form like(22,M),(22,F),(33,F),(52,F). This method is following remaining attributes such as zip code and disease. This approach ensure, provide the security to the table.

III. PROPOSED ALGORITHM

A. *m-privacy*

Definition: Given an *n* set of records which is provided by set of providers *P* and the Cartesian product method is applied for all sensitive attributes. Then anonymization technique is applied for all sensitive data while publishing the data to the public. Let consider the T={t1,t2,t3,…t*n*} be set of records which are horizontally distributed among multiple data providers as P={P1,P2,P3,….P*n*}, such as that Ti E T is a set of records provided by Pi. Let Assume the as is the sensitive attribute with domain Ds. If the record has the more sensitive attributes then a newly obtained sensitive attribute it can be defined as Cartesian product of all sensitive attributes. Then Q is define as conjunction of privacy constraints: Q1^Q2^……^Qn. If T* satisfies Q, then it says Q (T*) =true.



T_1

Name	Age	Zip	Disease
Alice	24	98745	Cancer
Bob	35	12367	Asthma
Emily	22	98712	Asthma

T_2

Name	Age	Zip	Disease
Dorothy	38	98701	Cancer
Mark	37	12389	Flu
John	31	12399	Flu

T_3

Name	Age	Zip	Disease
Sara	20	12300	Epilepsy
Cecilia	39	98708	Flu

T_4

Name	Age	Zip	Disease
Olga	32	12337	Cancer
Frank	33	12388	Asthma

Table VI. Data provider

Table VI describes the m-privacy approach with an example data. Assume the hospital, which means data provider provide the data with a set of records such as T1, T2, T3, T4 as shown in the Table I. Then each record contains a quasi-identity attribute (Name, Age, Zip code as zip) and the sensitive attributes (Disease). And the privacy constraint Q is defined as Q=Q1^Q2, where Q1 is k-anonymity with k=3 and Q2 is *l*-diversity with *l*=2 .Then both anonymized table T*a and T*b satisfies Q. Example in the T1, T2, T3 and T4 tables are joined in one table then the value of the age attribute are sorted in that table. Then the table are spit a three bucket. In each bucket. The value of the age attribute has a constraints, such means interval values. And that the value of intervals is assigned to the each tuple in its correspondent bucket. In T^*a, the intervals of the age attribute are [20-30] for first bucket, [31-35] for second bucket and [36-40] for third bucket. And the values of the zip attribute are encrypted and the values of the sensitive attribute are collapsed, such as the first value of the sensitive attribute are taken the preference first and assigned to the first tuple it is one of privacy method, shown in T*a table. The notion of m-privacy, which inhibit data knowledge of an m-adversaries with respect to a given privacy constraints. From the Table VII, *T*b* is an anonymized data which satisfies *m*-privacy (*m* = 1) with respect to *k*-anonymity (k=3) and *l*-diversity (*l* = 2). The value of the age attribute has taken the same interval levels for all tuples and the value of the zip attribute are encrypted differently for different buckets. In previous linguistic it ensure the increase the privacy to data while published to the public.

IV. PSEUDO CODE

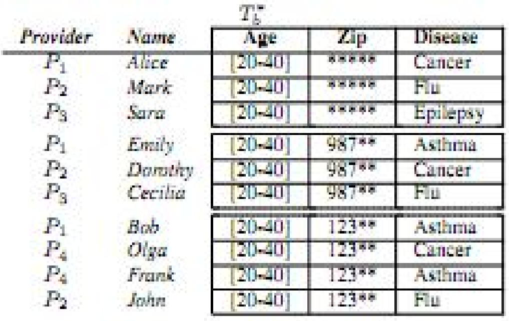
Algorithm 1: m-privacy algorithm

```
Z= {T}
While Z is not empty
 Remove the first B from Z=Z-{B}
Split B into two buckets B1 and B2
Set intervals as Interval= [Bi (1)-Bi (n)]
i= i+1;
Encrypt the QID
Encrypt the Sensitive data and interchange its position
Release to the public
```

V. EXPERIMENT RESULTS

A. *Formulization for m-privacy*

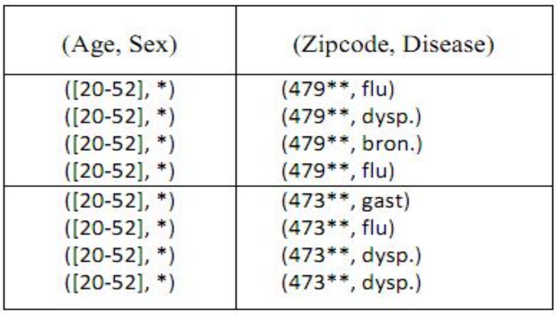
Let T be the set of data table which contains d attributes. a= {a1, a2…... an} and their domain attributes are {d [a1], d [a2]…d [an]}. A tuple t can be represented as t= (t [a1], t [a2]…... t [an]) where t [ai] is the value of ai of tuple t. Definition: An attribute partition consists of some subset of A. Which means each attribute belongs to exactly one subset. Hence each subset of attribute is called a column, let be a C1, C2.



T_5^*

Provider	*Name*	Age	Zip	Disease
P_1	*Alice*	[20-40]	*****	Cancer
P_2	*Mark*	[20-40]	*****	Flu
P_3	*Sara*	[20-40]	*****	Epilepsy
P_1	*Emily*	[20-40]	987**	Asthma
P_2	*Dorothy*	[20-40]	987**	Cancer
P_3	*Cecilia*	[20-40]	987**	Flu
P_1	*Bob*	[20-40]	123**	Asthma
P_4	*Olga*	[20-40]	123**	Cancer
P_4	*Frank*	[20-40]	123**	Asthma
P_2	*John*	[20-40]	123**	Flu

Table VII. m-privacy



(Age, Sex)	(Zipcode, Disease)
([20-52], *)	(479**, flu)
([20-52], *)	(479**, dysp.)
([20-52], *)	(479**, bron.)
([20-52], *)	(479**, flu)
([20-52], *)	(473**, gast)
([20-52], *)	(473**, flu)
([20-52], *)	(473**, dysp.)
([20-52], *)	(473**, dysp.)

Table VIII. Overlapping one attribute per column

Table VII illustrate the about the m-privacy methods. The original Table contains the original data, then first QID (Age) data are shorted with order and table is spliced into two bucket. After that the first value of Age attribute in first bucket is taken as first interval level and the last value of the age attribute in first bucket is taken as ending interval level and the that interval levels are applied for each tuple in Age attribute in first bucket. This method is applied for each bucket. Then the data of Sex attribute and the Zip Code are encrypted. The Table IV illustrated about the One attribute-per-column slicing the data of the Age attribute column is not interchanged but the data of the other QID column are interchanged its position, then Age attribute and sex attribute are merged and then the zip code and sex attribute are merged like that zip code . And sensitive attribute are merged and finally the data of the sensitive attribute column are anonymized. Overlapped sliced table obtained by overlapping the attributes in the Table V & VII. The attributes in the table 5 are replaced with the m-privacy Table VII. It shows better data utility than the existing anonymization techniques.

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose a new approach called slicing with the m-privacy technique to privacy-preserving microdata publishing. Slicing overcomes the limitations of generalization and bucketization and preserves better utility while protecting against privacy threats. We illustrate how to use slicing to prevent attribute disclosure and membership disclosure. Our experiments show that slicing preserves better data utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute. In future it can formed with three attribute per column with overlapping strategy.

REFERENCES.

1. Slawomir Goryczka, Benjamin, C. Xiong, LI. Fung,M., "m-privacy for Collaborative Data Publishing" {IEEE} Trans. Knowledge Data Eng., volume: pp no:99 year 2013.

2. Tiancheng, Li. Ninghui, Li. Ian Molloy and Jian Zhang, "slicing: a new approach for privacy preserving data publishing", {IEEE} Trans. Knowledge Data Eng., vol.23, no.2, 2011.

3. Mohammed, N. Fung, B. C. M. Hung, P. C. K. and Lee, C., "Centralized and distributed anonymization for high-dimensional healthcare data," ACM Trans. on Knowl. Discovery from Data, vol. 4, no. 4, pp. 18:1–18:33, October 2010.

4. Jiang, W. Clifton, C. "Privacy-preserving distributed k-anonymity," in DBSec, vol. 3654, 2005, pp. 924–924.

5. Jiang, W. Clifton, C., "A secure distributed framework for achieving k-anonymity," VLDB J., vol. 15, no. 4, pp. 316–333, 2006.

6. Lindell, Y. Pinkas, B., "Secure multiparty computation for privacy-preserving data mining," The Journal of Privacy and Confidentiality, vol. 1, no. 1, pp. 59–98, 2009.

7. Burke, R. Mobasher, B. Zabicki, R. Bhaumik, R., "Identifying attack models for secure recommendation," in Beyond Personalization: A Workshop on the Next Generation of Recommender Systems, 2005.
8. Sweeney, L., "Uniqueness of Simple Demographics in the U.S. Population," Carnegie Mellon University, Tech. Rep., 2000.
9. Gal, T.S. Chen, Z. Gangopadhyay, A., "A privacy protection model for patient data with multiple sensitive attributes," IJISP, vol. 2, no. 3, pp. 28–44, 2008.
10. Li, T. Li, N., "t-Closeness: Privacy beyond k-anonymity and l-diversity," in ICDE, 2007.