ISSN: 2321-7782 (Online)
Volume 2, Issue 11, November 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Privacy Preserving With Slicing Technique

Ratna Raju Mallela[1]
M.Tech Student
Department of Information Technology
V.R Siddhartha Engineering College
Andhra Pradesh – India

T. Lakshmi Surekha[2]
Assistant Professor
Department of Information Technology
V.R Siddhartha Engineering College
Andhra Pradesh – India

Abstract: ***Privacy preserving data publishing (PPDP) provides methods for publishing colleting and stored information. For providing safety so many privacy methods have been designed. In this paper give a brief description of several anonymous techniques for privacy preserving micro data publishing. Several anonymization techniques are designed for privacy preserving microdata publishing. Such Anonymization techniques are Generalization and Bucketization. But in generalization losses huge amount of information for high dimensional data. Bucketization doesn't prevent the identity disclosure and have a clear separation between sensitive and quasi-identifier attribute. To overcome those drawbacks introduce a new technique called slicing. Which segmentation the data both vertically and horizontally. Slicing can be used to prevent identification disclosure, and the main advantage of slicing is handling high dimensional data or large volume of information. Our experimental results shows that slicing is better than generalization for data utility, and providing privacy preserving than bucketization.***

Keywords: Privacy preservation, Anonymization, Data publishing, Security, Slicing.

I. INTRODUCTION

Data mining that is sometimes also experienced as Knowledge Discovery Data (KDD) is the work on of analyzing data from different perspectives and adding it into utilitarian information. Data mining is the taking out the fraught selective information from the large data sets such as data warehouse, Micro data sprains back records each of which contains selective selective information about an mortal entity. Many microdata anonymization proficiencies have been proposed and the most democratic ones are generalization [1],[2] with k-anonymity [2] and bucketization [3] with l diversity [4]. For privacy in Microdata publishing a novel proficiency called slicing is used that the sectionalizations the data both horizontally and vertically.

Slicing maintains better data usefulness than generalization and can be used for membership revelation protection. It can wield high dimensional data [1]. A securer system is required that can that can with stand high dimensional data handling and sensitive attribute revelation failures. These quasi–identifiers are set of attributes are those that in compounding can be linked with the external information to remainder. These are three grades of attributes in microdata. In the case of both anonymization proficiencies, first identifiers are moved out from the data and then segmentations the tuple's into buckets.

In generalization, transubstantiates the quasi-identifying values in each bucket into less particular and semantically constant so that tuple's in the same bucket cannot be differentiated by their QI values. One splits up the SA values from the QI values by randomly transposing the SA values in the bucket in the bucketization. The anonymized data consist of a set of buckets with transposing sensitive attribute values. Existing works mainly views datasets with a single sensitive attribute while patient data consists multiple sensitive attributes such as diagnosis and operation.

Data slicing can also be used to prevent membership revelation [11] and is efficient for high dimensional data and preserves better data usefulness. We introduce a novel data anonymization proficiency called slicing to meliorate the current state of the

art. Data has been zoned horizontally and vertically by the slicing. Vertical partitioning is done by pigeonholing attributes into columns based on the correlations among the attributes. Horizontal segmentation is done by pigeonholing tuple's into buckets. Slicing preserves usefulness because it groups highly correlated attributes together and upholds the correlations between such attributes. When the data set contains QIs and one SA, bucketization has to split up their correlation. Slicing can group some QI attributes with the SA for upholding attribute correlations [5] with the sensitive attribute. In this paper we introduce to develop efficient Tuple partition algorithm for privacy preserving in each user stipulation present in our data sets. In this standards of developing application is better and effectual solution for privacy of each user process. In this theatrical performance of the data set present in data base which assets efficient and squeezed data process. Our experimental results give good processing of the security retainers in recent applications of each user history process.

II. Related Work

2.1 Data Collection and Data Publishing

A typical premise of data collection and publishing is reported. In the data collection phase the data holder assembles data from record proprietors. As shown in the fig.1 data-publishing phase the data bearer waivers the collected data to a data miner or the public who will then transmit data mining on the brought out data.

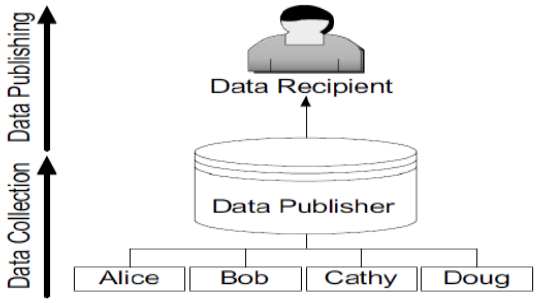

Fig. 1 Collection and Publishing of data

2.2 Privacy-Preserving Data Publishing

The privacy-preserving data publishing has the to the highest degree basic form that data holder has a table of the form: D (Identifier, Quasi Identifier, Sensitive Attributes, non-Sensitive Attributes) containing information that identifies record proprietors. Quasi Identifier is a set of attributes that could possible identify record owners. Sensitive Attributes consist of sensitive pertinently information. Non-Sensitive Attributes contains all attributes that do not fall into the premature three categories.

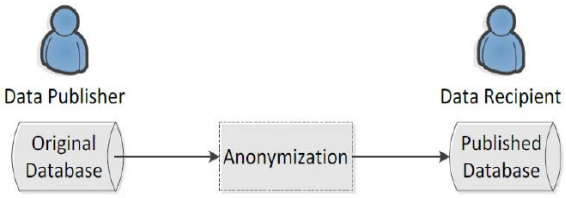

Fig. 2 Model of PPDP

2.3 Data Anonymization

Data Anonymization is a technology that transitions clear text into a non-human decipherable form. The technique for privacy-preserving data publishing has picked up a lot of attention in tardily years. Most democratic anonymization techniques are Generalization [1] and Bucketization [3]. The main difference between the two-anonymization proficiencies lies in that bucketization does not generalize the QI attributes.

2.4 Generalization

Generalization is one of the unremarkably anonymized approaches that supplant quasi-identifier values with values that are less specific but semantically reproducible. All quasi-identifier valuates in a group would be generalized to the stallion group extent in the QID space. If at least two proceedings in a group have distinct values in a certain column then all information about that item in the current group is turned a loss. QID used in this process admits all possible items in the lumber. In order for generalization to be efficacious, records in the same bucket must be tightlipped to each other so that generalizing the records would not fall behind too much information [1]. The data analyst has to make the uniform statistical distribution assumption that every value in a generalized interval/set is evenly possible to execute data analysis or data mining tasks on the generalized table [12]. This importantly trims down the data service program of the generalized data.

2.5 Bucketization

Bucketization is to sectionalization the tuple's in T into buckets and then to split up the sensitive attribute from the non-sensitive ones by every which way permuting the sensitive attribute values privileged each bucket.

We use bucketization as the method acting of manufacturing the published data from the original table T. We apply an sovereign random permutation to the column holding in S-values within each bucket [3]. The resulting set of buckets is then published. While bucketization has more salutary data utility than generalization it has several confinements [10]. Bucketization does not prevent membership revelation because bucketization publishes the QI values in their master copy forms. Bucketization requires a clear legal separation between QIs and SAs. In many data sets it is undecipherable which attributes are QIs and which are SAs. By ramifying the sensitive attribute from the QI attributes. Bucketization discontinues the attribute correlations between the QIs and the SAs. The anonymized data consist of a set of buckets with commuted sensitive attribute values [3]. Bucketization has been victimized for anonymizing high-dimensional data.

III. Basic idea of Data Slicing

Data slicing method sectionalizations the data both horizontally and vertically, which we discussed antecedently. The method sectionalization the data both horizontally and vertically [1]. This abbreviates the dimensionality of the data and preserves better data service program than bucketization and generalization.

Data slicing can be done in four stages:

- Partitioning attributes and columns

  An attribute partition dwells of several subsets of A that each attribute belongs to incisively one subset [6]. Consider only one sensitive attribute S one can either consider them singly or consider their joint distribution.

- Partitioning tuple's and buckets

  Each tuple conks to precisely one subset and the fixed of tuple's is called a bucket.

- Generalization of buckets

  A column generalization represents each value to the neighborhood in which the value is curbed.

- Matching the buckets

  We have to tally whether the buckets are matching. From each one tuple must be in one bucket only but not in many buckets.

Slicing:

The master copy microdata marched QI values and SI attributes. As shown in the Table I patient data in a hospital. Data consists of Age, Sex, Zip, desease.

TABLE I
ORIGINAL MICRODATA PUBLISHED.

Age	Sex	Zipcode	Disease
22	F	570004	Viral fever
24	M	570012	Flu
33	F	570061	Heart disease
52	F	625005	Cancer
55	M	625007	Dyspepsia
60	M	625110	Flu

The steganography that preserves the most information is “logarithmically”. The first tuple are screened out into buckets and then for each bucket because same attribute value may be generalized other than when they appear in different buckets.

TABLE II
GENERALIZED DATA

Age	Sex	Zipcode	Disease
22-50	*	570***	Viral fever
22-50	*	570***	Flu
22-50	*	570***	Heart disease
51-70	*	625***	Cancer
51-70	*	625***	Dyspepsia
51-70	*	625***	Flu

Table II testifies the generalized data of the viewed data in the above table. One column contains QI values and the other column comprises SA values in bucketization also attributes are districted into columns. In below table III we tincture the bucketization data. One singles out the QI and SA appraises by indiscriminately permuting the SA values in apiece bucket.

TABLE III BUCKETIZED DATA

Age	Sex	Zipcode	Disease
22	M	570004	Flu
22	F	570012	Heart disease
33	F	570061	Viral fever
52	F	625005	Dyspepsia
55	M	625007	Flu
60	M	625110	Cancer

The basic melodic theme of slicing is to belly-up the tie-up cross columns, to preserve the within each tie-up column. It represses the dimensionality of data and preserves better utility program. Data slicing can also palm high-dimensional data.

TABLE IV SLICED DATA

(Age, Sex)	(Zipcode, Disease)
(22, M) (22, F) (33, F)	(570061, Heart disease) (570004,Viral fever) (570012,Flu)

(52, F) (55,M) (60,M)	(625110,Flu) (625005,Cancer) (625007,Dyspepsia)

IV. Slicing Algorithm

We now exhibit an efficacious slicing algorithm to attain l-diverse slicing [4]. Our algorithm dwells of three phases: attribute partitioning, column generalization, and third one is tuple partitioning.

Attribute partitioning

This algorithm cleavage attributes so that extremely correlated attributes are in the like column. This is honest for both alternative and privacy. In terms of data alternative, pigeonholing extremely correlated attributes preserves the correlations among those attributes [6]. In monetary value of privacy, the association of uncorrelated attributes exhibits higher denomination takes chances than the association of highly correlated attributes because the associations of uncorrelated attribute values is much less sponsor and thus more diagnosable.

Column generalization

While away column generalization is not a requisite phase, it can be useful in versatile aspects. First, column generalization may be necessitated for idempotent disclosure protection. If a column value is unparalleled in a column [7], a tuple with this unparalleled column value can only have one touching bucket. This is not respectable for privacy protective covering, as in the encase of generalization/bucketization where each tuple can blend in to only one equivalence-class/bucket. The primary winding problem is that this unparalleled column value can be distinguishin. In this case, it would be useful to lend oneself column generalization to ensure that each column value seems with at least some relative frequency.

Second, when column generalization is lent oneself, to carry through the same level of privacy versus attribute disclosure, bucket sizings can be smaller. While column generalization may solvent in information loss, smaller bucket-sizes allow more upright data service program. Therefore, there is a trade-off amongst column generalization and tuple partitioning.

Tuple partitioning

The algorithm prolongs two data structures: one a lineup of buckets Q and second a set of sliced buckets SB. Ab initio, Q moderates only one bucket which admits all tuples and SB is vacuous [8]. For each curling, the algorithm murders a bucket from Q and dissevers the bucket into two buckets. If the sliced table after the split gratifies l-diversity, then the algorithm redacts the two buckets at the remnant of the queue Q Otherwise, we cannot fragmented the bucket anymore and the algorithm redacts the bucket into SB. When Q becomes vacuous, we have reckoned the sliced table. The set of sliced buckets is in SB.

V. Conclusion

Privacy preserving is the major undertaking in recent data mining diligence which particularizes processing operations in each user lay out in the data set monstrance. For serving this application process efficaciously, tralatitious we pose to develop Slicing with multi-dimensional data palming trading operations in each sectionalization present in the pinned down processing covering of each user. For unparalleled appellation unconscious process security of the each specifies and uprise commercial and modish technique. In this paper we project to develop Tuple partition algorithm for efficacious swearing out application outcomes which are arrogated to perform details of each with filtering conditions available in recent epoch application process of the specified data sets histrionics. Our experimental results show efficacious processing in unassailable format of the delineated field format present in the archetype data set deputation. Furthermore we advise to develop inscription schemas for processing efficient guarantor events in recent and recrudesced data sets.

VI. FUTURE WORK

We deliberate slicing where each attribute is in on the dot one column. Lengthiness is the whimsey of imbrication slicing, which replicates an attribute in more than one column.

Our experiments testify that random grouping is not very efficacious. We architectural plan to design more efficacious tuple grouping algorithms. Another steering is to design data mining undertakings using the anonymized data [9] computed by versatile Anonymization techniques.

Slicing protect privacy by divulging the association of uncorrelated attributes and preserve data utility program by preserving the association between extremely correlated attributes. Another fulfilling reinforcement of slicing is that it can palm high dimensional data.

References

1. P. Samarati, “Protecting Respondent’s Privacy in Microdata Release,” IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6, pp. 1010-1027, Nov./Dec. 2001.
2. L. Sweeney, “k-Anonymity: A Model for Protecting Privacy,” Int’l J. Uncertainty Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.
3. D.J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J.Y. Halpern, “Worst-Case Background Knowledge for Privacy- Preserving Data Publishing,” Proc. IEEE 23rd Int’l Conf. Data Eng. (ICDE), pp. 126-135, 2007.
4. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “l-Diversity: Privacy Beyond k-Anonymity,” Proc. Int’l Conf. Data Eng. (ICDE), p. 24, 2006.
5. H. Cramt’er, Mathematical Methods of Statistics. Princeton Univ. Press, 1948.
6. L. Kaufman and P. Rousueeuw, “Finding Groups in Data: An Introduction to Cluster Analysis,” John Wiley & Sons, 1990.
7. X. Xiao and Y. Tao, “Anatomy: Simple and Effective Privacy Preservation,” Proc. Int’l Conf. Very Large Data Bases (VLDB), pp. 139-150, 2006.
8. K. LeFevre, D. DeWitt, and R. Ramakrishnan, “Mondrian Multidimensional k-Anonymity,” Proc. Int’l Conf. Data Eng. (ICDE), p. 25, 2006.
9. C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating Noise to Sensitivity in Private Data Analysis,” Proc. Theory of Cryptography Conf. (TCC), pp. 265-284, 2006.
10. N. Koudas, D. Srivastava, T. Yu, and Q. Zhang, “Aggregate Query Answering on Anonymized Tables,” Proc. IEEE 23rd Int’l Conf. Data Eng. (ICDE), pp. 116-125, 2007.
11. M.E. Nergiz, M. Atzori, and C. Clifton, “Hiding the Presence of Individuals from Shared Databases,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD), pp. 665-676, 2007.
12. C. Aggarwal, “On k-Anonymity and the Curse of Dimensionality,” Proc. Int’l Conf. Very Large Data Bases (VLDB), pp. 901-909,2005.

