|| Volume 2 || Issue 5 || November 2017||ISSN (Online) 2456-0774

INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH AND ENGINEERING TRENDS

Privacy Preserving Data Publishing Using Slicing

Dinesh D. Jagtap, Prof. R. A. Auti

Student, CSE Department, Everest Educational Society's Group Of Institutions Aurangabad ,Maharashtra, India[1]

Asst. Professor, CSE Department, Everest Educational Society's Group of Institutions Aurangabad Maharashtra, India[2]

dinesh.jagtapd@gmail.com

Abstract— **There are number of techniques are researched and proposed for anonymization of data for privacy preserving data mining Data anonymization techniques for privacy – preserving data publishing has got the more importance in recent year. In any large organization such as hospitals, government agency, private firms there is a huge micro data is generated on daily basis so, maintaining the privacy and identify of any entity such like as patient, bank customer is having the highest priority. Anonymity is the way which the can hide anyone's identity or make it concealed. Most common and well known generalization & bucketization are the two techniques which are design for privacy preserving micro data publishing. The previous study of generalization shows that it loses subsequent amount of data while performing generalization process. Bucketization, on the other hand not applicable for preventing member ship disclosure. In this paper, we proposed an innovative technique called slicing, using this we can partitions the data both horizontally as well as vertically. In this paper we also shown that this technique preserves improved data utility than generalization technique and it also support membership disclosure protection.**

Keywords:- *Privacy Preserving, Membership Disclosure, Data Anonymization, Generalization, Bucketization*, Tuples, sensitive attributes (SAS)

I INTRODUCTION

Publishing of detail data by maintaining privacy has been studied widely in last couple of years. Data anonymization is a technique that converts a clear text into a non-human readable form. Most popular Privacy Preserving techniques are generalization & bucketization. In both methods attribute values are mainly classified into three categories.(i) Identifier by which we can uniquely identify any individual such as name or social security number.(ii) Some may be Quasi- identifier (QI) such as zip-code, age and gender whose values when combined taken together can possibly identify and individual.(iii) Some attributes are subtle or sensitive attributes, which are unknown to the opponent and considered as sensitive, like Disease and salary. In both methods first identifiers are truncated from the data and then tuples will be partitioned into buckets.

II EXISTING SYSTEM

From recent studies it is observe that first generalization losses substantial amount of data particularly for high dimensional data. Generalization based privacy preserving is not suitable for high-dimensional data. To make generalization more operative records belongs to the same bucket must be closed to each other so that performing generalization of records does not result into any kind of information loss. With comparison to bucketization, it provides efficient data utilization over generalization, but it is also having several limitations. The reason is bucketization issues the I values as it is in their actual forms so an challenger can easily discovers whether an individual is present in published data or not.

III PROPOSED SYSTEM

Here we are presenting a novel data anonymity technique called overlapping slicing. This technique divides the provided data set horizontally and vertically. In Vertical partitioning mechanism by attributes are grouped into column on the basis of attributes correlation among other attributes. And horizontally partitioning is performed by tuples grouping into bucket. And finally within each bucket each column values are randomly. Her column linking with different column is breaked but their association between each column is maintained. The algorithm of privacy preserving mainly contains the three steps attribute portioning column generalization and tuple partitioning

IV VARIOUS ANONYMIZATION TECHNIQUES

A. Generalization

Performing anonymization using generalization is one of the basic idea. In this approach the QI values are put in place of values that are less specific but semantically constant. So, all QI identifier values from one group are generalized to the whole cluster extent in the QID space. If minimum two relations in one group contains dissimilar values in a other column (i.e. one is having item and the other don't), in this case all information related to that item will get lost.. All possible items from the log are grouped into the QID. Quasi-identifier may be present in large number, And possible items may be in the order of thousands so making generalization of that would acquire tremendously high information loss, And makes the data unusable. To make the generalization more efficient, records from the same bucket must be close to each other so making generalization would not result into data loss. In the case of multi- dimensional data, most data points are having the same distance between each other. To initiate data analysis process of the table after generalization, the data analyst should assume that every value in generalized set is equally probable. So he has to make uniform distribution of all the values.

B. Bucketization

In this technique tuple T is partitioned into bucket and then non – sensitive values are get separated from sensitive attribute values within each bucket. Then the resultant data consists of the permuted sensitive values. In simple way let us set the partitions of bucketization more formally. First of tuples are get partitioned into buckets (partition the T horizontally), and inside every bucket random variation are applied to the column which contain s-values. After this generated set of buckets B is published. It is also having some limitations .Membership disclosure is not prevented in bucketization. Because of here QI values are published in their original forms so it is easy for adversary to find out whether published data contains record for certain user or not. Second it requires clear segregation between sensitive values and quassi identifiers.

C. Slicing

Slicing is a newly invented data anonymization technique. This technique divides the provided data set horizontally and vertically. In Vertical partitioning mechanism by attributes are grouped into column on the basis of attributes correlation among other attributes. And horizontally partitioning is performed by tuples grouping into bucket. And finally within each bucket each column values are randomly. here column linking with different column is braked but their association between each column is maintained. The algorithm of privacy preserving mainly contains the three steps attribute portioning column generalization and tuple partitioning. Slicing results into an efficient data utility because grouping of highly correlated data attributes is done and correlations between such attributes is maintained. So, slicing algorithm mainly classified into three steps.

1. Partitioning of attribute values.
2. Generalization of column.
3. Partitioned the tuples.

V SYSTEM ARCHITECTURE

A. Data Collection And Data Publishing

Figure 1 Describes the typical scenario for of collecting and publishing the data, the data holder obtains data from record owners i.e. Alice and Bob as shown in figure. Then in the data publishing phase, the collected data is forwarded to a data beneficiary, who will further perform data mining operation on the published data.

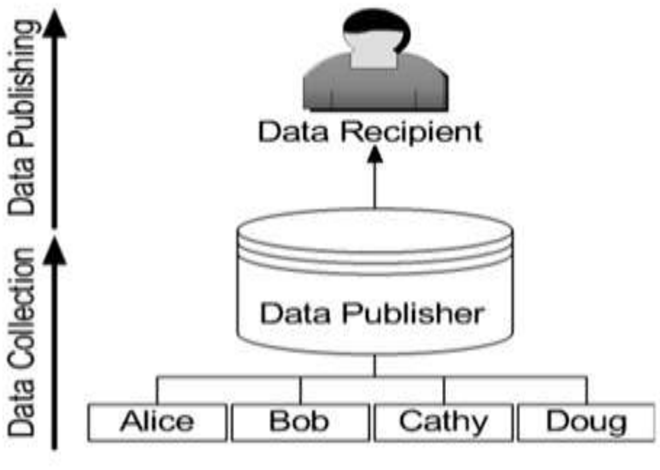

Figure 1 Collecting & Publishing data.

B Privacy-Preserving Data Publishing:

In this process of Privacy-Preserving Data Publishing (PPDP), the data can contains the set of attributes categorized into Clear Identifier, Quasi Identifier, Sensitive Attributes and non-Sensitive Attributes. Where Explicit Identifier is having attribute set, which individually identify any person such as name or social security number. When the values of quassi-identifier is taken into consideration combinely it, can potentially identify and individual. Some attributes are may be sensitive attributes (SAS), which are anonymous to the adversary and considered, sensitive, such as Disease and salary. And Non-Sensitive Attributes contains rest of the attributes which does not belongs to above mentioned three categories. These four set of attributes are considered as disjointly. Maximum studies adopt that every record in the table signifies a dissimilar record owner.

C Data Anonymization

In this phase clear text is converted into human non-readable form. Data anonymization technique for PPDP is gaining a lot much interest from last few years. Micro-data contains information about a household, a person, or an organization. Most popular Privacy Preserving techniques are generalization & bucketization. In both methods attributes are classified into three categories.(i) Identifier which can uniquely identify and individual such as name or social security number.(ii) Some may be Quasi- identifier (QI) such as zip code age and sex whose values when taken together can potentially identify and individual.(iii) Some attributes are sensitive attributes (SAS), which are unknown to the adversary and considered, sensitive, such as Disease and salary. In both methods first. Of all identifiers are removed from the data. The benefit of data anonymization is that using it we can transfer the information between two parties like two persons, departments, agencies, by overcoming threat factor of unintended disclosure, and in confident environments that empowers estimation and analytics post-anonymization.

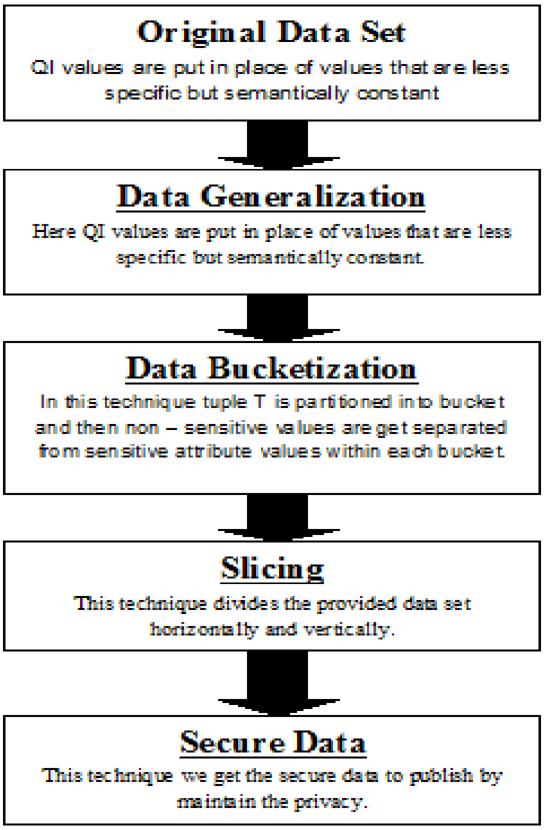
Original Data Set
QI values are put in place of values that are less specific but semantically constant
Data Generalization
Here QI values are put in place of values that are less specific but semantically constant
Data Bucketization
In this technique tuple T is partitioned into bucket and then non – sensitive values are get separated from sensitive attribute values within each bucket
Slicing
This technique divides the provided data set horizontally and vertically.
Secure Data
This technique we get the secure data to publish by maintain the privacy.

Figure 2: Proposed System Architecture

VI ALGORITHMS USED

Bucketization, generalization and such another algorithms having main focus on maintaining privacy, however they cannot prevent the attribute disclosure. So to solve this problem slicing algorithm is used. Slicing algorithm is mainly having three steps:

1. Attribute partitioning
2. Column generalization
3. Tuple partitioning

We will now describe these three phases.

A. Attribute Partitioning

As its name suggests, this algorithm makes partitions of attributes so, attributes which are highly correlated are placed into the similar column. To maintain utility and privacy this mechanism is somewhat good in nature. To achieve data utility, highly interrelated attributes are assembled and the associations among those attributes are maintained. In terms of privacy, the association of uncorrelated attributes is having higher risk of identification over the highly correlated attributes association. Because uncorrelated attribute associations is frequent and thus more specific to recognize.

B. Column Generalization

Generalization of column is not an essential phase, it can be helpful in some aspects. To maintain the membership disclosure protection column generalization is required. If the value of column value is unique in a column, then a tuple having this unique column value can only belongs to one matching bucket. In terms of maintaining privacy protection as in the case of generalization and bucketization where each tuple can only have single matching buckets. The main trouble is that the values which are unique in column can be identify. It would be beneficial to make column generalization mechanism to make sure that every columns value seems with at least some frequency. Second, to accomplish the equal privacy against attribute disclosure after performing column generalization, the size of bucket can be smaller. While column generalization may result in data loss, slighter bucket-sizes permit efficient utility of data. Therefore, the mechanism of tuple partitioning and column generalization is the alternative for each other.

C. Tuple Partitioning

Slicing algorithm is having two set of data structures:1) a queue containing the group of buckets Q 2) and second set contains the sliced buckets SB. Initially, Q contains a buckets having all the tuples and SB is empty. For every repetition, the algorithm eliminates a bucket from the set Q and again divides the bucket into two more buckets. If the l-diversity is satisfied by sliced table, then the algorithm places the two buckets at the last of queue Q. Otherwise, bucket cannot be split anymore and the algorithm puts the bucket into SB. When group of bucket Q becomes empty, then we got the set of computed sliced table. SB is the group sliced buckets.

VII FUTURE SCOPE AND CONCLUSION

The limitations of generalization and bucketization are somewhat reduced by slicing approach. And it also maintains efficient data utility while protecting against privacy related issues. Attribute disclosure and membership disclosure is prevented by slicing mechanism. Slicing gives better result in terms of efficient data utility over generalization and is also more advanced and powerful than bucketization in workloads having the set of sensitive attribute. We assume slicing where every attribute is in just one column. An extension is the notion of overlapping slicing, which duplicates an attribute in more than one column. The random grouping is not the perfect solution to maintain privacy. We proposed to design more operative algorithm for tuple grouping. Additional way is to plan a data mining tasks using the anonymzed data generated by various Anonymizations methods. Privacy is preserve in slicing by breaking the association between attributes which are highly correlated. And reserve data utility by maintaining the highly correlated attribute association. Another important benefit of slicing is it can handle large and high-dimensional amount of data. The simple idea suggested by this work is first of all consider the characteristic of data before performing annonymization on it, after analysis of characteristics performs anonymization strategy.

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]C.Aggarwal. On k-anonymity and the curse of dimensionality. In VLDB, pages 901–909, 2005.

[3]Asuncion and D. Newman. UCI machine learning repository, 2007.

[4]Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the sulq framework. In PODS, pages 128–138, 2005.

[5]J.Brickell and V. Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In KDD, pages 70–78, 2008.

[6]B.-C. Chen, R. Ramakrishnan, and K. LeFevre. Privacy skyline: Privacy with multidimensional adversarial knowledge. In VLDB, pages 770–781, 2007.

[7]H. Cramt’er. Mathematical Methods of Statistics. Princeton, 1948.

[8]Dinur and K. Nissim. Revealing information while preserving privacy. In PODS, pages 202–210,2003.

[9] A. Machanavajjhala, D. Kifer, J. Gehrke,and M. Venkitasubramaniam. “L-diversity: Privacy beyond k-anonymity”. InICDE 2006.