

SLICING: A SECURED DESIGN FOR THE MICRO DATA PUBLICATION

Theyyagura M K Reddy¹, B. Venkata Suresh Reddy²

¹M.Tech (IT) Scholar, ²Assistant Professor

Nalanda Institute of Engg and Tech. (NIET), Siddharth Nagar, Guntur, A.P, (India)

ABSTRACT

Anonymization of data is always a fascinating thing for the researchers in last few years. Preservation of the privacy while moving for the data mining in a data warehouse is the research subject which is said to be increasing the attention of the research community. Data publishing by preserving the privacy of the data has received a lot of attention and also described as the main problem that is said to be making a challenge of hoe to secure a database of high dimension. When we consider a big organisation where large amount of confidential data is available, then in such a situation where that much amount of confidential data is present and which should be secured from the accessing of unauthorized users. Such a security to the confidential data is needed since it may be misused if it is accessed by an unauthorized user for a variety of purposes. In order to mitigate these concerns, a number of techniques are being proposed which are intended to perform the data mining in a way where it can also preserve the privacy of the confidential data. Several anonymization techniques are available namely the Generalization and Bucketization which are designed for the purpose of preserving the data privacy of the publishing of micro data. Though these are proposed for the privacy preservation we found that when a very high dimension data is to be mined and we follow the principle of Generalization then that may lose some information and on the other hand the Buketization which does not prevent the membership disclosure. We here present another methodology of anonymization technique named Slicing. Usage of Slicing have such a significance is that it is able of handling a very high dimensional data. Slicing is able of preserving better data utility when compared to generalization and also able of preventing the membership disclosure. Here we manly focus on effective method that is able of providing better data utility and also can handle high dimensional data.

I. INTRODUCTION

We come across many situations where we might choose to withhold our identity. This situation is mostly seen in the act of charity where the benefactors do not wish to be acknowledged. When a person feels threatened will try to alleviate that threat by following anonymity. When we consider some situations remaining anonymous will be illegal. There are “Stop and identify” statutes in nearly 24 states of United States which requires persons who are detained to self-Identify when a law enforcement officer requests. In the past few years due to increase in ability of storing the personal data about the users and the increase in the sophistication of the data mining algorithms that can leverage this personal information made the problem of privacy preserving data mining become more important. Here we have a lot of anonymization techniques that have been researched in order to perform data mining by privacy preserving. The technique of data anonymization for privacy preserving has got

a lot of attention these years. Micro data which is also called as the detailed data contains information about a person, an organisation or a house hold. The most famous among the anonymization techniques are Generalization and Bucketization. We may find a number of attributes in each record that can be categorized as

- 1) Identifiers like name or an Identification number that be used to identify a person uniquely
- 2) Some of the attributes may be sensitive such as disease he is suffering from and salary
- 3) Some of the identifiers are called as the quasi identifiers like zip code, age and sex. If all of these values if taken together have the ability of potentially identify an individual.

II. BACKGROUND

Anonymity is fined as the condition where one's name or identity is concealed or unknown. This will serves very valuable social purposes and are able to empower the individuals against the institutions by surveillance limitations but on the other hand the same phenomenon can be used by wrong doers also to hide their actions or to avoid the accountability of their deeds and gives then an ability to allow anonymous access to services which will avoid the tracking of user behaviour such as user location and user's personal information frequency of the service usage by the user etc. If a file transfer is done then there may be some information about the origin of the file will be known. Here the information of the sender is known or traced by the data that is logged after the file is sent.

2.1 Security vs. Anonymity

The technique of anonymity is considered as a very powerful technique of privacy protection. The design of the internet which is stateless and decentralized is the one which is particularly suitable for anonymous behaviour. Though anonymous actions can ensure privacy they are not supposed to be used as the only source of ensuring the privacy as they also allow harmful activities such as slander, spamming and other harmful activities without a fear of reprisal. Security is the main concern that says that one should be able of detecting and catching the individuals who are conducting illegal behaviour such as conspiring for terrorist acts, hacking and conducting a fraud activity. Lawful needs for privacy should be allowed and at the same time the ability for conducting the harmful anonymous behaviour without repercussions and responsibility by saying the name of privacy should not.

2.2 Anonymity vs. Privacy

Anonymity and privacy are not the same. There is a distinction between anonymity and privacy is clearly seen in the information technology context. Anonymity defines about being able f sending the contents of the email in plain text which is in easily readable format but it will not give any information which gives the reader a chance to trace who the writer of that mail is. On the other hand privacy defines as the ability of sending the email to a person in an encrypted and unreadable manner instead sending as plain text. Here by the email knowing about the person who wrote the email may be possible but reading the contents of the mail without his authorization is not possible. Anonymity plays a very crucial role where the identity of the author of the message is said to be the main concern and on the other hand privacy plays a very important role when the contents of the message are the primary concern.

III. PROPOSED WORK

3.1 Problem Statement

The privacy of the Database in other words the data that is stored and which is said to be confidential is important to private citizens and organizations alike. Sometimes security against theft of the storage systems involving servers, hard drives, laptops and desktops can be given by private professionals. While coming for the case of organisations they should ensure that all the storage management interfaces and backups of database which may be at on site or off site must maintain their integrity. If there is an attack on a database then it is the responsibility of the organisation to take the defensive measures against those attacks. This will first limit the classification of the data that is done according to the importance of the data immediately. After doing so the encryption methods can be employed this may help in protecting databases and applications on their sensitivity levels. But we know that the best method of protecting the privacy of the database is the prevention of attacks on it. One of the privacy preserving methods of the database will include a mechanism of assessing the database regularly to check whether any exploitation is done or not and signing that it has been compromised. If the organisation succeeds in detecting the exploit or the indications of database compromising before that attack actually happens in real then there will be a chance for rectification of the database with only a little and reversible damage which will mostly saves the data and time for the organisation.

3.2 Goals

One of the most important research problems is for handling the high dimensional data. As we discussed above the privacy Preservation for high dimensional database is important issue of concern. We do have to popular techniques of data anonymization they are namely Generalization and Bucketization which are used for making the data anonymization of high dimensional data sets. Here the proposed mechanism of slicing is able of preserving better data utility than generalisation and can be used for membership disclosure procedure also.

IV. SLICING ALGORITHMS

Partitioning of the attributes in to columns is the first thing to do under our proposed mechanism. A subset of attributes will be present in each and every column. This will make the partition of the table vertically. Slicing will also make buckets by partitioning the tuples. Here each bucket will be having a subset of tuples.

4.1 Attribute Partition and Columns

There are several subsets in the partition of attributes, where a rule exists as each attribute belongs to exactly one subset. Here a column is a subset of Attributes. For simplicity of our discussion here we consider only one sensitive attribute S. But if a condition arises where the data contains multiple attributes which are to be treated sensitive then we can either consider all those as a joint distribution or separately. When in the situation where there is no loss of generality let us consider that the column C that contains the sensitive attribute S be the last column. This column since it contains the attribute that is treated as the sensitive attribute it will be called as the sensitive column.

Here the algorithm will partition the attributes so that the attributes that are highly correlated will fall in same column. This will work well for both data utility and data privacy. When we consider the terms of data utility the grouping of highly correlated attributes preserves the correlations among the attributes that are correlated. When we consider privacy the association of uncorrelated data attributes showed higher identification risks

when compared with the correlated data attributes. This is because of the fact that the uncorrelated attributes are less frequent when compared with the correlated attributes which make them easy to identify. So in order to protect the privacy of uncorrelated attributes it is better to break the relations between them. Here in this phase we first compute the correlation between the pairs of attributes and then cluster attributes based on their attributes.

4.2 Column Generalization

Here we discuss about the second phase where the tuples will be generalised to satisfy minimal frequency requirement. Here we will point out that that column generalization which is not an indispensable phase present in our algorithm. Here we have seen that the concept of Bucketization which is proposed for the privacy preservation of data provides the same level of privacy preservation to that of generalisation. Although we do not need column generalization as a phase it will help us in several aspects. For identity/membership disclosure protection the column generalisation is very much important. Here a unique column can have only one matching bucket when it has a tuple present in it. But it is not good for providing the privacy preservation as that is present in the case of Bucketization/generalisation where as we discussed each tuple will belong to only one equivalence-class/bucket. Here we do have a main problem and that is the unique column can be easily identified. Here in this case applying of column generalization will be useful for ensuring that each value will appear with at least some frequency. Here when we apply the column generalisation for achieving the same level of privacy against attribute disclosure and the bucket size can be smaller. While we go with the column generalisation it may result in information loss, so here we prefer the smaller bucket sizes since they will allow better data utility. Because of this there is a trade-off between the column generalisation and tuple partition. It will serve as the subject for future work. The existing anonymization techniques can be used for column generalisation. These algorithms can be applied only on the attributes that are suitable and contain only one column to ensure the anonymity requirement.

4.3 Tuple Partitioning

In this phase the tuples are partitioned into buckets. And here no generalization is applied to the tuples. Here this algorithm maintains two data structures. They are listed below

- 1) a queue of buckets Q
- 2) a set of sliced buckets SB

Here initially the queue of buckets Q contains only one bucket which will include all the tuples and keeping the set of sliced buckets SB empty. Then after each iteration the algorithm removes a bucket from Q and splits that single bucket into two buckets. Here if the l-diversity is satisfied with the slice table after the split then the two buckets are kept at the end of the queue Q.

Otherwise the bucket cannot be split anymore and the algorithm will keep it in SB. We have computed the SB when the Q is empty.

V. EXPERIMENTS

Here there is an important research problem and that is handling the high dimensional data. As we discussed privacy preserving of high dimensional database is important. The two popular data anonymization techniques are Generalization and Bucketization. The aim of these two techniques is privacy preserving and publishing of micro data. Here we are proposing a new technique named slicing which is better than both the generalization

and Bucketization for preserving the privacy in high dimensional data sets mainly when relational databases are taken into consideration. Our proposed technique is able of preserving data utility in a better way than that of generalization and also can be used for protection of membership disclosure. The existing data anonymization techniques can be classified into several dimensions and they are:

i. Nature of Data

This is for dealing with the techniques that are proposed for a tabular data which will represent the information about entities, their quasi-identifiers (or) confidential data sets and their sensitive information. Here the item set data which is consider as representation of transactional data which is as associating with the people with the set of items that are purchased in a transaction and a graph data which is supposed as a representation of sensitive associations between the entities.

ii. Anonymization approaches

A variety of approaches are used in the proposed anonymization technique which include

- a) Suppression, where we find that some information is removed from the data
- b) Generalization, where the information is made into sets.
- c) Perturbation, is used where noise is added to the data and
- d) Permutation, is used when sensitive associations between entities are to be swapped.
- e) Anonymization Objectives:

By ensuring the published data various privacy goals are achieved and they have certain properties. Here these properties are mostly the k-anonymity where each individual in the database and it must be indistinguishable from k-1 others. Then l-diversity is to be considered which seeks for ensuring the sufficient diversity among the sensitive information that are associated with the individuals and finally the other goals which will aim to prevent certain interferences which n bare based on assumptions that are based upon the knowledge be held by the attacker.

Here we say that the person must be aware of various data anonymization technique. And also they also should be aware of the relational databases and how a better privacy preserving can be given to those records which are available in these relational database tables.

Here we are facing a huge problem and that is none other than the Privacy issue of the database. In many of the government and private organisations where we consider the Hospitals, various multinational companies, colleges and so on. Where there exist large databases available. For such databases privacy should be maintained properly.

VI. DISCUSSIONS AND FUTUREWORK

Our proposed paper is depicted with careful anonymization, which can provide strong and robust methods for privacy protection for the data of the individuals which is in published or shared databases and at the same time not losing much utility of the data. When we consider the Privacy protection anonymity is considered as a very powerful technique. Here a new approach for privacy preserving called Slicing is being proposed. Our proposed mechanism Slicing is a promising technique when we consider the scenario of handling high-dimensional data. By using the proposed system named slicing we will be able of hiding the data from the real world. This is achieved so because the identity of the records will be either changed or removed and only then it is shown to the real world. This will help us in to ways those are making the database more secure and the same time will

keep the data privacy. Here our comparison of the proposed mechanism with the existing system proved that it is better than Generalization and Bucketization which are considered to be the efficient mechanisms before our proposed mechanism.

Here in this paper the mechanisms are based on three dimensions:

- 1) Here we design an intuitive, simple and robust privacy model
- 2) Here we design an efficient anonymization technique is used that works with the real world databases
- 3) Here we develop a framework which will evaluate privacy and utility tradeoff.



VII. CONCLUSION

Here we proposed a novel approach of which serves us a better way of Privacy preservation for the data where we consider a very high dimensional data in the relational databases. We casually produce the relations among support standards of a container. This might misplace statistics usefulness. Through separating qualities into supports, we defend confidentiality through contravention the connotation of uncorrelated qualities and reservation statistics usefulness by conserving the reminder among extremely connected features.

REFERENCES

- [1] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the sulq framework. In PODS, pages 128–138, 2005.
- [1] H. Cramt'er. Mathematical Methods of Statistics. Princeton, 1948.
- [2] A. Asuncion and D. Newman. UCI machine learning repository, 2007.
- [3] M. E. Nergiz, M. Atzori, and C. Clifton. Hiding the presence of individuals from shared databases. In SIGMOD, pages 665–676, 2007.

AUTHOR PROFILE

	Theyyagura MK Reddy is currently pursuing M.Tech in the Department of Information Technology, from Nalanda Institute of Engineering & Technology (NIET), Siddharth Nagar, Kantepudi(V), Sattenapalli(M), Guntur (D), Andhra Pradesh, Affiliated to JNTU-KAKINADA.
	B. Venkata Suresh Reddy working as Assistant Professor at Nalanda Institute of Engineering & Technology (NIET), Siddharth Nagar, Kantepudi(V), Sattenapalli (M), Guntur (D), Andhra Pradesh, Affiliated to JNTU-KAKINADA.