IJDCST @October Issue- V-1, I-6, SW-47
ISSN-2320-7884 (Online)
ISSN-2321-0257 (Print)

State of Art Algorithm for High Dimensional Data using Slicing

[1]Vuyyuru Asha, [2]M.Sailaja , [3]V.Siva Parvathi ,

[1] Student, PVPSIT, KANURU, VIJAYAWADA, KRISHNA DIST.

[2] Assistant Prof, PVPSIT,KANURU,VIJAYAWADA,KRISHNA DIST.

[3] Assistant Prof, PVPSIT,KANURU,VIJAYAWADA,KRISHNA DIST.

Abstract: Privacy-preserving data mining is the area of data mining that used to safeguard sensitive information from unsanctioned disclosure. Many of them have recognized the potential value of these data as an information source for making business decisions. Privacy-preserving data publishing has seen rapid advances that have lead to an increase in the capability to store and record personal data about consumers and individuals. Maintain the privacy for the high dimensional database has become important aspect. The problem of privacy-preserving data mining has become more important in recent years because of the increasing ability to store personal data about users. Privacy-preserving data publishing (PPDP) provides methods and tools for publishing useful information while preserving data privacy. Number of techniques such as randomization and kanonymity, bucketization, generlization has been proposed in recent years in order to perform privacy-preserving data mining. Recent work has shown that generalization loses considerable amount of information, especially for high-dimensional data. In this paper, we provide state-of-art methods for privacy for the high dimensional databases, and focus on effective method that can be used for providing better data utility and can handle high-dimensional data.

I. INTRODUCTION

In recent years, due to increase in ability to store personal data about users and the increasing sophistication of data mining algorithms to leverage this information the problem of privacy-preserving data mining has become more important. No. of anonymization techniques have been researched in order to perform privacy-preserving data mining. Data Mining which is sometimes also called as Knowledge Discovery Data (KDD) is the process of analyzing data from different perspectives and summarizing it into useful information.

Data mining is used by many companies with a strong consumer focus such as financial, marketing organizations, communication and retail. Extraction of hidden predictive information from large databases is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Most of existing work is formulated in the following context: Several organizations publish detailed data (also called *micro data*) about individuals for research or statistical purposes.

Sensitive personal information may be disclosed in this process, due to the existence in the data of quasi-identifying attributes or simply quasi-identifiers (QID). An attacker can join the QID with external information to reidentify individual records. Previous privacy-preserving techniques focus on anonym zing personal data that have a fixed schema with a small number of dimensions.

The big challenge for companies today – particularly as email and the Internet make sharing and distributing corporate information easier than ever - is to strike the right balance between providing workers with appropriate access and protecting sensitive information as much as possible. As companies continue to consolidate databases and streamline operations to maximize efficiency and the protection of data from external threats, this user- and role-based security model no longer complies with "need-to-know" security best-practices.

a. **Data Collection and Data Publishing**

As shown in the fig.1 the typical scenario of the data collection and the data publishing is viewed. In the phase of the data collection, the data holder collects the data from the records owner (likely consider Miller and Jackson). Where as in the phase of the data publishing, the data holder releases the collected data to a data miner or the public (i.e. Data Recipient), who will then conduct data mining on the published data.

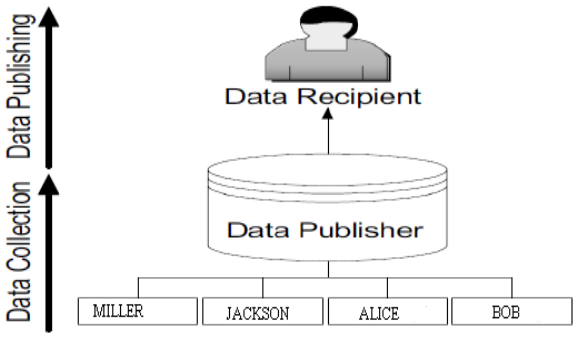

Figure 1: Data collection and Data Publishing

b. **Privacy-Preserving Data Publishing**

In the most basic form of privacy-preserving data publishing (PPDP) the data holder has a table of the form: D (Explicit Identifier, Quasi Identifier, Sensitive Attributes, non-Sensitive Attributes)

Explicit Identifier: It containing information that explicitly identifies record owners

Quasi Identifier: It is a set of attributes that could potentially identify record owners

Sensitive Attributes: It consist of sensitive person-specific information

Non-Sensitive Attributes: It contains all attributes that do not fall into the previous three categories

Most works assume that each record in the table represents a distinct record owner.

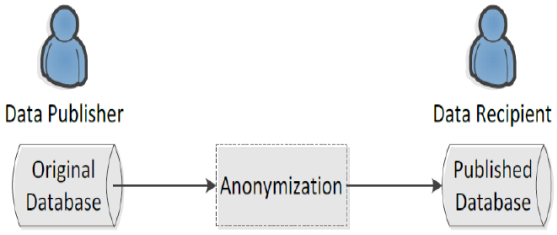

Figure.2: A Simple Model of PPDP

II. BACKGROUND

Two main Privacy preserving paradigms have been established:

- *k-anonymity:* It prevents identification of individual records in the data
- *l-diversity:* It prevents the association of an individual record with a sensitive attribute value

K-anonymity:

The database is said to be K-anonymous where attributes are suppressed or generalized until each row is identical with at least k-1 other rows. K-Anonymity guarantees that the data released is accurate. The K-anonymity proposal focuses on two techniques in particular:

a. Generalization
b. Suppression

To protect respondents' identity when releasing micro data, data holders often remove or encrypt explicit identifiers. De-identifying data provide no guarantee of anonymity. Released information often contains other data that can be linked to publicly available information to re-identify respondents and to infer information that was not intended for release. K-anonymity demands that every tuple in the microdata table released be indistinguishably related to no fewer than k respondents. k-anonymity with protection techniques that preserve the truthfulness of the data.
One definition of privacy which has come a long way in the public arena and is accepted today by both legislators and corporations is that of k-anonymity. The guarantee given by k-anonymity is that no information can be linked to groups of less than k individuals.

l- Diversity:

Consider we have a group of k different records that all share a particular quasi-identifier. It's good enough, in that an attacker cannot identify the individual based on the quasi-identifier. The distribution of target values within a group is referred to as "l-diversity". There exist two broad categories of l-diversity techniques: generalization and permutation-based. existing generalization method would partition the data into disjoint groups of transactions are well represented sensitive items, such that each group contains sufficient records with l-distinct.

III. STATE-OF-THE-ART TECHNOLOGIES

The state-of-the-art implementations of cloud computing is presented. Technologies used for cloud computing are describes here.

Architectural Design of Data Centre:

A **data center** is a facility used to house computer systems and associated components like telecommunications and storage systems.

Key Design Areas

- ✓ Resilience - ensuring maximum uptime without compromising on performance.
- ✓ Availability - business continuity is especially important.
- ✓ Performance - the faster the better; in a predictable manner.
- ✓ Security - ensuring data separation and controlled access to all Data centre resources.
- ✓ Effective architecture for data separation – a common infrastructure that provides facilities for network-based backup and efficient back-end network access.
- ✓ Predictable Failover - for maximum service availability.

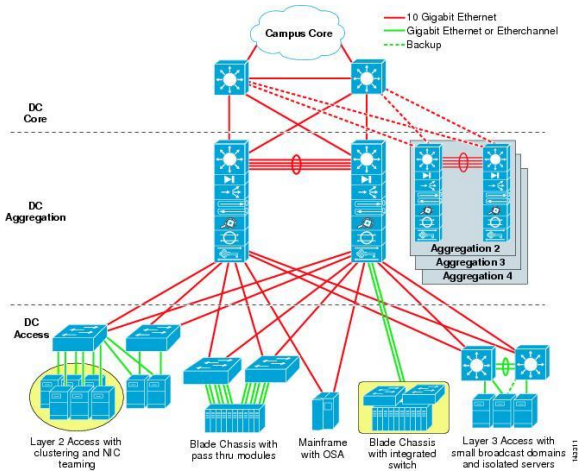

Figure 1: Data Centre Design

Distributed File System over Cloud:

We focused on Google File System that is a proprietary distributed file system developed by Google and specially designed to provide efficient. GFS is designed and optimized to run on data centers to provide extremely high data throughputs, survive individual server failures and low latency.

Distributed Application Framework over Clouds:

MapReduce is a software framework introduced by Google to support distributed computing on large data sets on clusters of computers. It consists of one Master to which client applications submit MapReduce jobs. Open source Hadoop MapReduce project is inspired by Google's work. Master pushes work out to available task nodes in the data centre striving to keep the tasks as close to the data as possible. Many organizations are using Hadoop MapReduce to run large data intensive computations.

IV. VARIOUS ANONYMIZATION TECHNIQUES

The main difference between the two anonymization techniques lies in that bucketization does not generalize the QI attributes.

Generalization:

Generalization is one of the commonly anonymized approaches that replaces quasi-identifier values with values that are less-specific but semantically consistent. All quasi-identifier values in a group would be generalized to the entire group extent in the

QID space. If at least two transactions in a group have distinct values in a certain column then all information about that item in the current group is lost. QID used in this process includes all possible items in the log. In order for generalization to be effective records in the same bucket must be close to each other so that generalizing the records would not lose too much information. In high-dimensional data most data points have similar distances with each other. The data analyst has to make the uniform distribution assumption that every value in a generalized interval/set is equally possible. This significantly reduces the data utility of the generalized data.

Bucketization:

First we term bucketization is to partition the tuples in T into buckets and then to separate the sensitive attribute from the non-sensitive ones by randomly permuting the sensitive attribute values within each bucket. We use bucketization as the method of constructing the published data from the original table T. Partitition the tuples into buckets and within each bucket we apply an independent random permutation to the column containing S-values. While bucketization has better data utility than generalization. bucketization does not prevent membership disclosure because bucketization publishes the QI values in their original forms. Bucketization requires a clear separation between QIs and SAs. In many data sets it is unclear which attributes are QIs and which are SAs.

V. PROBLEM STATEMENT

When people speak about database privacy then they usually are referring to the protection of information contained within digital databases and of the databases themselves. Database privacy is a concept that is important to organizations and private citizens alike. Organizations have the responsibility to protect clients' information because their clients entrust them to do so. No. of steps that organizations can take to help safeguard databases and the data they hold. Assigning proper authentication levels to database workers providing unique authentication credentials for each application. Privacy professionals also can secure storage systems against theft involving desktops, off-site, servers. Organizations should ensure that storage management interfaces and all database backups maintain their integrity. It is an organization's responsibility to take defensive measures. This might first entail the immediate classification of data according to importance. One method of database privacy protection might include assessing a database regularly for exploits and signs that it has been compromised. An organization can detect exploits or indications of database compromising before the threat becomes real and unmanageable.

VI. PROPOSED WORK

As Privacy-preservation for the high dimensional database has become important in many ways. Database of any organization, medical, company‟s is a confidential database. We introduce a data anonymization technique called slicing to improve the current state of the art. Slicing partitions the data set both vertically and horizontally. The 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. Within each bucket values in each column are randomly permutated to break the linking between different columns. This reduces the dimensionality of the data and preserves better utility than generalization and bucketization. Slicing preserves utility because it groups highly correlated attributes together, and preserves the correlations between such attributes. Note that when the data set contains QIs and one SA, the bucketization has to break their correlation; slicing. The key intuition that slicing provides privacy protection is that the slicing process ensures that for any tuple. Slicing first partitions attributes into columns. It also partition tuples into buckets. This horizontally partitions the table. Values in each column are randomly permutated to break the linking between different columns.

VII. CONCLUSION

An important research problem is for handling high-dimensional data. Privacy Preservation for high

dimensional database is important. Two popular data anonymization technique Generalization and Bucketization are used. These techniques are designed for privacy preserving microdata publishing. Proposed work include a slicing technique which is better than generalization and bucketization for the high dimension data sets. It preserves better data utility than generalization and can be used for membership disclosure protection. And one more important advantage of slicing is that it can handle high-dimensional data.

VIII. REFERENCE's

[1] Neha V. Mogre, Prof. Girish Agarwal, Prof. Pragati Patil, "Privacy Preserving Data Publishing Concepts and Techniques" , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013.

[2] D. Mohanapriya , Dr. T.MeyyappanOn Slicing Technique For Privacy Preserving Data Publishing, *International Journal of Computer Trends and Technology (IJCTT) – volume 4 Issue 5–May 2013*.

[3] Amar Paul Singh, Dhanshri Parihar. A Review of Privacy Preserving Data Publishing Technique. *International Journal of Emerging Research in Management &Technology*, *ISSN: 2278-9359 (Volume-2, Issue-6), june-2013*.

[4] Benjamin C. M. Fung, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu, "Privacy Preserving Data Publishing Concepts and Techniques" ,*Data mining and knowledge discovery series (2010).*

[5] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati On K-Anonymity. In Springer US, Advances in Information Security (2007).

[6] Latanya Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):557–570, 2002.

[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.