ISSN 2278 - 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 1, MARCH 2012

A survey on Privacy Preserving Data Mining

R.Natarajan[1], Dr.R.Sugumar, M.Mahendran, K.Anbazhagan

Research Scholar, Dept. of CSE, CMJ University, Shillong, Meghalaya, India

Associate Professor, Dept. of CSE, VelMultitech Dr. RR Dr.SR Engineering College, Chennai, India

Research Scholar, Dept. of CSE, CMJ University, Shillong, Meghalaya, India

Research Scholar, Dept. of CSE, CMJ University, Shillong, Meghalaya, India

Abstract: Privacy Preserving Data Mining (PPDM) is a research area concerned with the privacy driven from personally identifiable information when considered for data mining. Therefore, PPDM has become an increasingly important field of research. PPDM is a novel research direction in data mining. A number of methods and techniques have been developed for privacy preserving data mining. This paper provides a complete review on PPDM and different techniques such as data partition, data modification, data restriction technique which could be used to prevent the data access from unauthorized users.

Keywords: preserving privacy, data modification, data restriction, data ownership

I. INTRODUCTION

The scope of information technologies and the internet in the past two decades has brought a wealth of individual information into the hands of commercial companies and government agencies. As hardware costs go down, organizations and it easier than ever to keep any piece of information acquired from the ongoing activities of their clients. Data owners constantly seek to make better use of the data they possess, and utilize data mining tools to extract useful knowledge and patterns from the data.

Privacy Preserving Data Mining (PPDM) is a research area concerned with the privacy driven from personally identifiable information when considered for data mining. Therefore, PPDM has become an increasingly important field of research. PPDM is a novel research direction in data mining. A number of methods and techniques have been developed for privacy preserving data mining (Philip S. Yu et al. 2010).

The set of criteria has been identified based on which a PPDM algorithm can be evaluated (Charu C. Aggarwal et al. 2008).

- Privacy level
- Hiding failure
- Data quality
- Complexity

The major challenges of PPDM method for association rule hiding are high information loss, expensive, difficult to recover original data after hiding and should be efficient enough for very large datasets (Wei Zhao et. al 2007).

PPDM is a research area concerned with the privacy driven from personally identifiable information when considered for data mining. This work addresses the privacy problem by considering the privacy and algorithmic requirements simultaneously. The objective of this work is to implement a distortion algorithm using association rule hiding for privacy preserving data mining which would be efficient in providing confidentiality and improve the performance (Charu C. Aggarwal et al. 2008).

The debate on PPDM has received special attention as data mining has been widely adopted by public and private organizations. We have witnessed three major landmarks that characterize the progress and success of this new research area: *the conceptive landmark*, *the deployment landmark*, and *the prospective landmark*. We describe these landmarks as follows:

The Conceptive landmark characterizes the period in which central figures in the community, such as O'Leary (1991, 1995), Piatetsky-Shapiro (1995), and others (Klösgen, 1995; Clifton & Marks, 1996), investigated the success of knowledge discovery and some of the important areas where it can conflict with privacy concerns. The key finding was that knowledge discovery can open new threats to informational privacy and information security if not done or

ISSN 2278 - 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 1, MARCH 2012

used properly. *The Deployment landmark* is the current period in which an increasing number of PPDM techniques have been developed and have been published in refereed conferences. The information available today is spread over countless papers and conference proceedings. The results achieved in the last years are promising and suggest that PPDM will achieve the goals that have been set for it.

The Prospective landmark is a new period in which directed efforts toward standardization occur. At this stage, there is no consent about what privacy preservation means in data mining. In addition, there is no consensus on privacy principles, policies, and requirements as a foundation for the development and deployment of new PPDM techniques. The excessive number of techniques is leading to confusion among developers, practitioners, and others interested in this technology. One of the most important challenges in PPDM now is to establish the groundwork for further research and development in this area.

A.Privacy Violation in Data Mining

Understanding privacy in data mining requires understanding how privacy can be violated and the possible means for preventing privacy violation. In general, one major factor contributes to privacy violation in data mining: the misuse of data.

Users' privacy can be violated in different ways and with different intentions. Although data mining can be extremely valuable in many applications (e.g., business, medical analysis, etc), it can also, in the absence of adequate safeguards, violate informational privacy. Privacy can be violated if personal data are used for other purposes subsequent to the original transaction between an individual and an organization when the information was collected (Culnan, 1993).

One of the sources of privacy violation is called data magnets (Rezgui et al., 2003). Data magnets are techniques and tools used to collect personal data. Examples of data magnets include explicitly collecting information through on-line registration, identifying users through IP addresses, software downloads that require registration, and indirectly collecting information for secondary usage. In many cases, users may or may not be aware that information is being collected or do not know how that information is collected. In particular, collected personal data can be used for secondary usage largely beyond the users' control and privacy laws. This scenario has led to an uncontrollable privacy violation not because of data mining itself, but fundamentally because of the misuse of data.

B.Defining Privacy for Data Mining

In general, privacy preservation occurs in two major dimensions: users' personal information and information concerning their collective activity. We refer to the former as individual privacy preservation and the latter as collective privacy preservation, which is related to corporate privacy in (Clifton et al., 2002).

- *Individual privacy preservation*: The primary goal of data privacy is the protection of personally identifiable information. In general, information is considered personally identifiable if it can be linked, directly or indirectly, to an individual person. Thus, when personal data are subjected to mining, the attribute values associated with individuals are private and must be protected from disclosure. Miners are then able to learn from global models rather than from the characteristics of a particular individual.
- *Collective privacy preservation*: Protecting personal data may not be enough. Sometimes, we may need to protect against learning sensitive knowledge representing the activities of a group. We refer to the protection of sensitive knowledge as collective privacy preservation. The goal here is quite similar to that one for statistical databases, in which security control mechanisms provide aggregate information about groups (population) and, at the same time, prevent disclosure of confidential information about individuals. However, unlike as is the case for statistical databases, another objective of collective privacy preservation is to protect sensitive knowledge that can provide competitive advantage in the business world.

In the case of collective privacy preservation, organizations have to cope with some interesting conflicts. For instance, when personal information undergoes analysis processes that produce new facts about users' shopping patterns, hobbies, or preferences, these facts could be used in recommender systems to predict or affect their future shopping patterns. In general, this scenario is beneficial to both users and organizations. However, when organizations share data in a collaborative project, the goal is not only to protect personally identifiable information but also sensitive knowledge represented by some strategic patterns.

II. Characterizing Scenarios of Privacy Preservation on the Web

- In this section, we describe two real-life motivating examples in which PPDM poses different constraints:
- *Scenario 1*: Suppose we have a server and many clients in which each client has a set of sold items (e.g., books, movies, etc.). The clients want the server to gather statistical information about associations among items in order to provide recommendations to the clients. However, the clients do not want the server to know some strategic patterns (also called sensitive association rules). In this context, the clients represent

ISSN 2278 - 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 1, MARCH 2012

companies and the server is a recommendation system for an e-commerce application, for example, fruit of the clients collaboration. In the absence of rating, which is used in collaborative filtering for automatic recommendation building, association rules can be effectively used to build models for on-line recommendation. When a client sends its frequent itemsets or association rules to the server, it must protect the sensitive itemsets according to some specific policies. The server then gathers statistical information from the non-sensitive itemsets and recovers from them the actual associations. How can these companies benefit from such collaboration by sharing association rules while preserving some sensitive association rules?

- *Scenario 2*: Two organizations, an Internet marketing company and an on-line retail company, have datasets with different attributes for a common set of individuals. These organizations decide to share their data for clustering to and the optimal customer targets so as to maximize return on investments. How can these organizations learn about their clusters using each other's data without learning anything about the attribute values of each other?
- Note that the above scenarios describe different privacy preservation problems. Each scenario poses a set of challenges. For instance, scenario 1 is a typical example of collective privacy preservation, while scenario 2 refers to individual's privacy preservation.

III. A TAXONOMY OF EXISTING PPDM TECHNIQUES

In this section, we classify the existing PPDM techniques in the literature into four major categories: data partitioning, data modification, data restriction, and data ownership as can be seen in Figure 1.

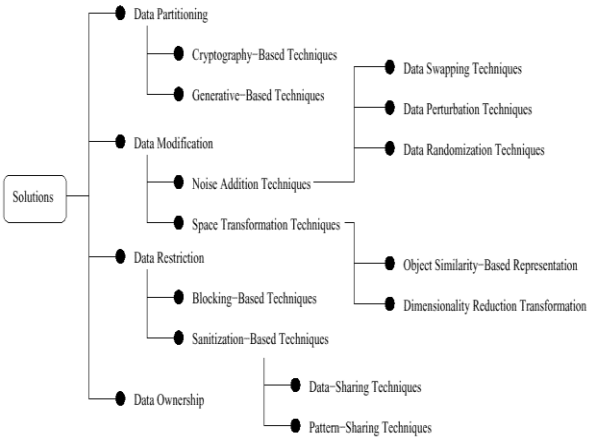


Figure 1. A taxonomy of PPDM techniques

A.Data Partitioning Techniques

Data partitioning techniques have been applied to some scenarios in which the databases available for mining are distributed across a number of sites, with each site only willing to share data mining results, not the source data. In these cases, the data are distributed either horizontally or vertically. In a horizontal partition, different entities are described with the same schema in all partitions, while in a vertical partition the attributes of the same entities are split across the partitions. The existing solutions can be classified into Cryptography-Based Techniques and Generative-Based Techniques.

- *Cryptography-Based Techniques*: In the context of PPDM over distributed data, cryptography-based techniques have been developed to solve problem of the following nature: two or more parties want to conduct a computation based on their private inputs. The issue here is how to conduct such a computation so that no party knows anything except its own input and the results. This problem is referred to as the Secure Multi-Party Computation (SMC) problem (Goldreich, Micali, & Wigderson, 1987). The technique proposed in (Lindell & Pinkas, 2000) address privacy-preserving classification, while the techniques proposed in (Kantarcioğlu & Clifton, 2002; Vaidya & Clifton, 2002) address privacy-preserving association rule mining, and the technique in (Vaidya & Clifton, 2003) addresses privacy-preserving clustering.
- *Generative-Based Techniques*: These techniques are designed to perform distributed mining tasks. In this approach, each party shares just a small portion of its local model which is used to construct the global model. The existing solutions are built over horizontally partitioned data. The solution presented in (Veloso et al., 2003) addresses privacy-preserving frequent itemsets in distributed databases, whereas the solution in

ISSN 2278 - 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 1, MARCH 2012

(Meregu & Ghosh, 2003) addresses privacy-preserving distributed clustering using generative models.

B.Data Modification Techniques

Data modification techniques modify the original values of a database that needs to be shared, and in doing so, privacy preservation is ensured. The transformed database is made available for mining and must meet privacy requirements without losing the benefit of mining. In general, data modification techniques aim at finding an appropriate balance between privacy preservation and knowledge disclosure. Methods for data modification include noise addition techniques and space transformation techniques.

- *Noise Addition Techniques*: The idea behind noise addition techniques for PPDM is that some noise (e.g., information not present in a particular tuple or transaction) is added to the original data to prevent the identification of confidential information relating to a particular individual. In other cases, noise is added to confidential attributes by randomly shuffling the attribute values to prevent the discovery of some patterns that are not supposed to be discovered. We categorize noise addition techniques into three groups: (1) data swapping techniques that interchange the values of individual records in a database (Estivill-Castro & Brankovic, 1999); (2) data distortion techniques that perturb the data to preserve privacy, and the distorted data maintain the general distribution of the original data (Agrawal & Srikant, 2000); and (3) data randomization techniques which allow one to perform the discovery of general patterns in a database with error bound, while protecting individual values. Like data swapping and data distortion techniques, randomization techniques are designed to find a good compromise between privacy protection and knowledge discovery (Evfimievski et al., 2002; Rizvi & Haritsa, 2002; Zang, Wang, & Zhao, 2004).
- *Space Transformation Techniques*: These techniques are specifically designed to address privacy-preserving clustering. These techniques are designed to protect the underlying data values subjected to clustering without jeopardizing the similarity between objects under analysis. Thus, a space transformation technique must not only meet privacy requirements but also guarantee valid clustering results. We categorize space transformation techniques into two major groups: (1) object similarity-based representation relies on the idea behind the similarity between objects, i.e., a data owner could share some data for clustering analysis by simply computing the dissimilarity matrix (matrix of distances) between the objects and then sharing such a matrix with a third party. Many clustering algorithms in the literature operate on a dissimilarity matrix (Han & Kamber, 2001). This solution is simple to be implemented and is secure, but requires a high communication cost (Oliveira & Zaïane, 2004); (2) dimensionality reduction-based transformation can be used to address privacy-preserving clustering when the attributes of objects are available either in a central repository or vertically partitioned across many sites. By reducing the dimensionality of a dataset to a sufficiently small value, one can find a trade-off between privacy, communication cost, and accuracy. Once the dimensionality of a database is reduced, the released database preserves (or slightly modifies) the distances between data points. In tandem with the benefit of preserving the similarity between data points, this solution protects individuals' privacy since the attribute values of the objects in the transformed data are completely different from those in the original data (Oliveira & Zaïane, 2004).

C.Data Restriction Techniques

Data restriction techniques focus on limiting the access to mining results through either generalization or suppression of information (e.g., items in transactions, attributes in relations), or even by blocking the access to some patterns that are not supposed to be discovered. Such techniques can be divided into two groups: Blocking-based techniques and Sanitization-based techniques.

- *Blocking-Based Techniques*: These techniques aim at hiding some sensitive information when data are shared for mining. The private information includes sensitive association rules and classification rules that must remain private. Before releasing the data for mining, data owners must consider how much information can be inferred or calculated from large databases, and must look for ways to minimize the leakage of such information. In general, blocking-based techniques are feasible to recover patterns less frequent than originally since sensitive information is either suppressed or replaced with unknowns to preserve privacy. The techniques in (Johnsten & Raghavan, 2001) address privacy preservation in classification, while the techniques in (Johnsten & Raghavan, 2002; Saygin, Verykios, & Clifton, 2001) address privacy-preserving association rule mining.
- *Sanitization-Based Techniques*: Unlike blocking-based techniques that hide sensitive information by replacing some items or attribute values with unknowns, sanitization-based techniques hide sensitive information by strategically suppressing some items in transactional databases, or even by generalizing information to preserve privacy in classification. These techniques can be categorized into two major groups: (1) data-sharing techniques in which the sanitization process acts on the data to remove or hide the group of sensitive association rules that contain sensitive knowledge. To do so, a small

ISSN 2278 - 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 1, MARCH 2012

number of transactions that contain the sensitive rules have to be modified by deleting one or more items from them or even adding some noise, i.e., new items not originally present in such transactions (Verykios et al., 2004; Dasseni et al., 2001; Oliveira & Zaïane, 2002, 2003a, 2003b); and (2) pattern-sharing techniques in which the sanitizing algorithm acts on the rules mined from a database, instead of the data itself. The existing solution removes all sensitive rules before the sharing process and blocks some inference channels (Oliveira, Zaïane, & Saygin, 2004). In the context of predictive modeling, a framework was proposed in (Iyengar, 2002) for preserving the anonymity of individuals or entities when data are shared or made publicly.

D.Data Ownership Techniques

Data ownership techniques can be applied to two different scenarios: (1) to protect the ownership of data by people about whom the data were collected (Felty & Matwin, 2002). The idea behind this approach is that a data owner may prevent the data from being used for some purposes and allow them to be used for other purposes. To accomplish that, this solution is based on encoding permissions on the use of data as theorems about programs that process and mine the data. Theorem proving techniques are then used to guarantee that these programs comply with the permissions; and (2) to identify the entity that receives confidential data when such data are shared or exchanged (Mucsi-Nagy & Matwin, 2004). When sharing or exchanging confidential data, this approach ensures that no one can read confidential data except the receiver(s). It can be used in different scenarios, such as statistical or research purposes, data mining, and on-line business-to-business (B2B) interactions.

Are These Techniques Applicable to Web Data?

After describing the existing PPDM techniques, we now move on to analyze which of these techniques are applicable to Web data. To do so, hereinafter we use the following notation:

- **WDT**: these techniques are designed essentially to support Web usage mining, i.e., the techniques address Web data applications only. We refer to these techniques as Web Data Techniques (WDT).

- **GPT**: these techniques can be used to support both public data release and Web-based applications. We refer to these techniques as General Purpose Techniques (GPT).

a) Cryptography-Based Techniques: these techniques can be used to support business collaboration on the Web. Scenario 2 (in Section: The Basis of Privacy-Preserving Data Mining) is a typical example of Web-based application which can be addressed by cryptography-based techniques. Other applications related to e-commerce can be found in (Srivastava et al., 2000; Kou & Yesha, 2000). Therefore, such techniques are classified as WDT.

b) Generative-Based Techniques: these techniques can be applied to scenarios in which the goal is to extract useful knowledge from large, distributed data repositories. In these scenarios, the data cannot be directly centralized or unified as a single file or database either due to legal, proprietary or technical restrictions. In general, generative-based techniques are designed to support distributed Web-based applications.

c) Noise Addition Techniques: these techniques can be categorized as GPT. For instance, data swapping and data distortion techniques are used for public data release, while data randomization could be used to build models for on-line recommendations (Zang et al., 2004). Scenario 1 (in Section: The Basis of Privacy-Preserving Data Mining) is a typical example of an on-line recommendation system.

d) Space Transformation Techniques: these are general purpose techniques (GPT). These techniques could be used to promote social benefits as well as to address applications on the Web (Oliveira & Zaïane, 2004). An example of social benefit occurs, for instance, when a hospital shares some data for research purposes (e.g., cluster of patients with the same diseases). Space transformation techniques can also be used when the data mining process is outsourced or even when the data are distributed across many sites.

e) Blocking-Based Techniques: in general, these techniques are applied to protect sensitive information in databases. They could be used to simulate an access control in a database in which some information is hidden from users who do not have the right to access it. However, these techniques can also be used to suppress confidential information before the release of data for mining. We classify such techniques as GPT.

f) Sanitization-Based Techniques: Like blocking-based techniques, sanitization-based techniques can be used by statistical offices who publish sanitized version of data (e.g., census problem). In
addition, sanitization-based techniques can be used to build models for on-line recommendations
as described in Scenario 1 (in Section: The Basis of Privacy-Preserving Data Mining).

g) Data Ownership Techniques: These techniques implement a mechanism enforcing data ownership by the individuals to whom the data belongs. When sharing confidential data, these techniques can also be used to ensure that no one can read confidential data except the receiver(s) that are authorized to do so. The most evident applications of such techniques are related to Web mining and on-line business-to-business (B2B) interactions.

Table 1 shows a summary of the PPDM techniques and their relationship with Web data applications.

ISSN 2278 - 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 1, MARCH 2012

PPDM Techniques Category	Category
Cryptography-Based Techniques	WDT
Generative-Based Techniques	WDT
Noise Addition Techniques	GPT
Space Transformation Technique	GPT
Blocking-Based Techniques	GPT
Sanitization-Based Techniques	GPT
Data Ownership Techniques	WDT

Table 1: A summary of the PPDM techniques and their relationship with Web data.

IV. REQUIREMENTS FOR TECHNICAL SOLUTIONS

A. Requirements for the development of technical solutions

Ideally, a technical solution for a PPDM scenario would enable us to enforce privacy safeguards and to control the sharing and use of personal data. However, such a solution raises some crucial questions:

- What levels of effectiveness are in fact technologically possible and what corresponding regulatory measures are needed to achieve these levels?
- What degrees of privacy and anonymity must be sacrificed to achieve valid data mining results?

These questions cannot have "yes-no" answers, but involve a range of technological possibilities and social choices. The worst response to such questions is to ignore them completely and not pursue the means by which we can eventually provide informed answers. The above questions can be to some extent addressed if we provide some key requirements to guide the development of technical solutions.

The following key words are used to specify the extent to which an item is a requirement for the development of technical solutions to address PPDM:

- **Must**: this word means that the item is an absolute requirement;
- **Should**: this word means that there may exist valid reasons not to treat this item as a requirement, but the full implications should be understood and the case carefully weighed before discarding this item.

a) Independence: A promising solution for the problem of PPDM, for any specific data mining task (e.g., association rules, clustering, and classification), should be independent of the mining task algorithm.

b) Accuracy: When it is possible, an effective solution should do better than a trade-off between privacy and accuracy on the disclosure of data mining results. Sometimes a trade-off must be found as in scenario 2 (in Section: The Basis of Privacy-Preserving Data Mining).

c) Privacy Level: This is also a fundamental requirement in PPDM. A technical solution must ensure that the mining process does not violate privacy up to a certain degree of security.

d) Attribute Heterogeneity: A technical solution for PPDM should handle heterogeneous attributes (e.g., categorical and numerical).

e) Communication Cost: When addressing data distributed across many sites, a technical solution should consider carefully issues of communication cost.

B.Requirements to guide the deployment of technical solutions

Information technology vendors in the near future will offer a variety of products which claim to help protect privacy in data mining. How can we evaluate and decide whether what is being offered is useful? The nonexistence of proper instruments to evaluate the usefulness and feasibility of a solution to address a PPDM scenario challenge us to identify the following requirements:

a) Privacy Identification: We should identify what information is private. Is the technical solution aiming at protecting individual privacy or collective privacy?

b) Privacy Standards: Does the technical solution comply with international instruments that state and enforce rules (e.g., principles and/or policies) for use of automated processing of private information?

c) Privacy Safeguards: Is it possible to record what has been done with private information and be transparent with individuals about whom the private information pertains?

d) Disclosure Limitation: Are there metrics to measure how much private information is disclosed? Since privacy has many meanings depending on the context, we may require a set of metrics to do so. What is most important is that we need to measure not only how much private information is disclosed, but also measure the impact of a technical solution on the data and on valid mining results.

e) Update Match: When a new technical solution is launched, two aspects should be considered: i) the solution should comply with existing privacy principles and policies; ii) in case of modifications to privacy principles and/or

ISSN 2278 - 1021

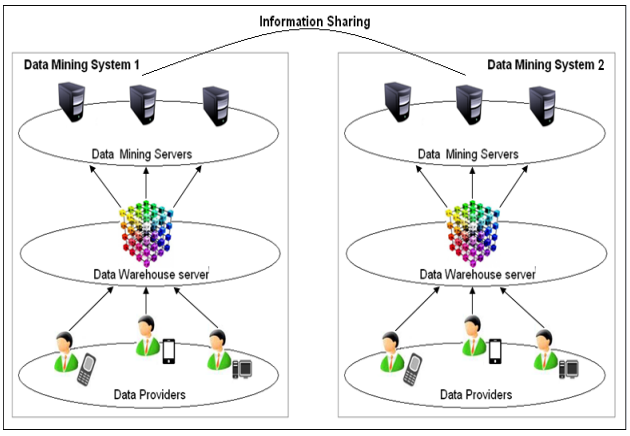
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 1, MARCH 2012

policies that guide the development of technical solutions, any release should consider these new modifications.

V. ARCHITECTURE OF PPDM

PPDM is usually carried out in multiple steps. First, the data being mined are collected from their sources, which are referred as data providers. In many systems, data providers are physically distributed, forming the bottom tier of the architecture of data mining systems, as shown in Figure 1.1. It shows privacy-preserving data mining usually has multiple steps that translate to a three-tiered architecture. The bottom tier has the data providers, the data owners, which are often physically distributed. The data providers submit their private data to the data warehouse server. This server, which constitutes the middle tier, supports online analytical data processing to facilitate data mining by translating raw data from the data providers into aggregate data that the data mining servers can more quickly process (Nan Zhang et. al 2007).

The data warehouse server stores the data collected in disciplined physical structures, such as a multidimensional data cube, and aggregates and recomputed the data in various forms, such as sum, average, max,and min. In an online survey system, for example, the survey respondents would be data providers who submit their data to the survey analyzer's data warehouse server; an aggregated data point might be the average age of all survey respondents. The aggregated data is more efficient to process than raw data from the providers. At the top tier are the data mining servers, which perform the actual data mining (Wei Zhao et. al 2010).



.

Figure 1.1 Basic architecture of Privacy-Preserving Data Mining

In a PPDM system, these servers do not have free access to all data in the data warehouse. In a hospital system, the accounting department can mine patients' financial data, for example, but cannot access patients' medical records. Developing and validating effective rules for the data mining server access to the data warehouse is an open research problem, besides constructing data mining models on its local data warehouse server, a data mining server might share information with data mining servers from other systems. The motivation for this sharing is to build data mining models that span systems. For example, several retail companies might opt to share their local data mining models on customer records to build a global data mining model about consumer behavior that would benefit all the companies. As Figure 1.1 shows, sharing occurs in the top tier, where each data mining server holds the data mining model of its own system. Thus, "sharing" means sharing local data mining models rather than raw data (Nan Zhang et. al 2007).

A.Based on the Improvement of Privacy Level

Privacy level offered by a privacy preserving technique, which indicates how closely the sensitive information, that has been hidden, can still be estimated. PPDM is a research strives to ensure that the privacy of each individual is maintained, yet present data mining results as accurately as possible. Data mining solutions that are privacy-aware should thus strive to provide highly accurate result sets while maintaining individual privacy.

From a philosophical point of view, Schoeman and Walters (2001) identify three possible definitions of privacy:

- Privacy as the right of a person to determine which personal information about himself or herself may be communicated to others.
- Privacy as the control over access to information about oneself.
- Privacy as limited access to a person and to all the features related to the person.

Clifton (2006) noted that since there are so many solutions in PPDM, it is difficult to simplify the research to the point that the solutions can be developed and implemented. Privacy-preserving data mining algorithms have been published in the research community in leading computing journals, yet the most obvious problem is that PPDM-enabled tools have not been widely adopted.

Preventing individual information disclosure has become increasingly important due to the number and size of data breaches during the last five years. There are countless examples of inadvertent disclosure of data. For example, in May, 2006, the Social Security numbers of

ISSN 2278 - 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 1, MARCH 2012

about 26.5 million U.S. veterans were stolen in a random burglary from a Veterans Affairs employee's house where a laptop was stolen (Torres et al 2007).

Xiong (2009) noted that with the increasing need to share data, protecting that data has also become important because sharing data with organizations in countries that have lesser privacy and security standards creates additional challenges. Organizations also put themselves at risk when they outsource their data processing activities to third-party vendors.

B.Based on Reducing the Hiding Failure

Hiding is the failure portion of sensitive information that is not hidden by the application of a privacy preservation technique.

Oliveira and Zaiane (2004) proposed a heuristic-based framework for preserving privacy in mining frequent itemsets. They focus on hiding a set of frequent patterns, containing highly sensitive knowledge. They propose a set of sanitized algorithms that only remove information from a transactional database, also known in the statistical disclosure control area as non-perturbative algorithms, unlike those algorithms, that modify the existing information by inserting noise into the data, referred to as perturbative algorithms. The first parameter is evaluated in terms of: Hiding Failure (ie) the percentage of restrictive patterns that are discovered from the sanitized database; Misses Cost (ie) the percentage of non-restrictive patterns that are hidden after the sanitization process; Artifactual Pattern, measured in terms of the percentage of discovered patterns that are artifacts.

The issue of k-anonymity is also important in the context of hiding identification in the context of distributed location based services. In this case, k-anonymity of the user-identity is maintained even when the location information is released. Such location information is often released when a user may send a message at any point from a given location (Bettini et. al 2006).

Zhong (2008) proposed an approach has been discussed for the case of horizontally partitioned data. This work discusses an extreme case in which each site is a customer which owns exactly one tuple from the data. It is assumed that the data record has both sensitive attributes and quasi-identifier attributes. The solution uses encryption on the sensitive attributes. The sensitive values can be decrypted only if therefore are at least k records with the same values on the quasi-identifiers. Thus, k-anonymity is maintained.

C.Based on the Improvement of Data Quality

Data quality after the application of a privacy preserving technique, considered both as the quality of data themselves and the quality of the data mining results after the hiding strategy is applied. In evaluating the data quality after the privacy preserving process, it can be useful to assess both the
quality of the data resulting from the PPDM process and the quality of the data mining results. The quality of the data themselves can be considered as a general measure evaluating the state of the individual items contained in the database after the enforcement of a privacy preserving technique.(Bertino et. al 2008).

In the scientific literature data quality is generally considered a multi-dimensional concept that in certain contexts involves both objective and subjective parameters. Among the various possible parameters, there are few most commonly used parameters exist namely,

- **Accuracy**: Measures the proximity of a sanitized value to the original value.
- **Completeness**: Evaluates the degree of missed data in the sanitized database.
    - **Consistency**: it is related to the internal constraints, that is, the relationships that must hold among different fields of a data item or among data items in a database (Wang et. al 2006).

D.Based on Reducing the Complexity

Complexity is the ability of a privacy preserving algorithm to execute with good performance in terms of all the resources implied by the algorithm. The complexity metric measures the efficiency and scalability of a PPDM algorithm. Efficiency indicates whether the algorithm can be executed with good performance, which is generally assessed in terms of space and time. Space requirements are assessed according to the amount of memory that must be allocated in order to implement the given algorithm (Bertino et. al 2008).

For the evaluation of time requirements, there are several approaches. The first approach is to evaluate the CPU time. Oliveira and Zaiane (2002) proposed a method to keep constant both the size of the database and the set of restrictive patterns, and then increase the size of the input data to measure the CPU time taken by their algorithm. An alternative approach would be to evaluate the time requirements in terms of the computational cost. In this case, it is obvious that an algorithm having a polynomial complexity is more efficient than another one with exponential complexity. Sometimes, the time requirements can even be evaluated by counting the average number of operations executed by a PPDM algorithm.

ISSN 2278 - 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 1, MARCH 2012

Kantarcioglu (2004) noted the performance is measured in terms of the number of encryption and decryption operations required by the specific algorithm. The last two measures, i.e. the computational cost and the average number of operations, do not provide an absolute measure, but they can be considered in order to perform a fast comparison among different algorithms.

In case of distributed algorithms, especially the cryptography-based algorithms, the time requirements can be evaluated in terms of communication cost during the exchange of information among secure processing. Specifically the communication cost is expressed as the number of messages exchanged among the sites that are required by the protocol for securely counting the frequency of each rule (Clifton et. al 2009).

VI. ISSUES IN DESIGNING A PPDM ALGORITHM

The major challenges that a PPDM algorithm for association rule hiding are information loss, expensive, recover original data after hiding and should be efficient enough for very large datasets (Agarwal et. al 2008).

1.Challenges of PPDM Algorithm

Information Loss: The information loss is defined as the ratio between the sum of the absolute errors made in computing the frequencies of the items from a sanitized database and the sum of all the frequencies of items in the original database. Inference control in databases, also known as Statistical Disclosure Control (SDC), is about protecting data so they can be published without revealing confidential information that can be linked to specific individuals among those to which the data correspond. This is an important application in several areas, such as official statistics, health statistics, e-commerce (sharing of consumer data), etc. Since data protection ultimately means data modification, the challenge for SDC is to achieve protection with minimum loss of the accuracy sought by database users (Aggarwal et al. 2008).

A.Expensive

Many of the protocols based on encryption use the idea introduced by Yao (2007). In Yao's protocol one of the parties compute a scrambled version of a boolean circuit for evaluating the desired function. The scrambled circuit consists of encryptions of all possible bit values on all possible wires in the circuit. The number of encryptions is approximately 4m, where m is the number of gates in the circuit. The encryptions can be symmetric key encryption, which has a typical ciphertext-length of 64 bits. The scrambled circuit is sent to the other party, which can then evaluate the circuit to get the final result. These approaches are, in general, expensive since they require complicated encryptions for each individual bit (Thomas b. Pedersen et. al 2007).

B.Recover original data after hiding

PPDM consists of number of techniques to retrieve the information from the large amount of database which consists of sensitive information also. k-anonymity is a method to suppress or generalize the data so that the data cannot be accessed by any unauthorized users. Once the data are suppressed or generalized using k-anonymity, it is very difficult to recover the original data (Agarwal et. al 2008).

C.Support of large datasets

Due to the continuous advances in hardware technology, large amounts of data can now be easily stored. Databases along with data warehouses today store and manage amounts of data which are increasingly large. For this reason, a PPDM algorithm has to be designed and implemented with the capability of handling huge datasets that may still keep growing. The less fast is the decrease in the efficiency of a PPDM algorithm for increasing data dimensions, the better is its scalability. Therefore, the scalability measure is very important in determining practical PPDM techniques (Bertino et. al 2008).

2. Requirements of a PPDM algorithm

A.Accuracy

The accuracy is closely related to the information loss resulting from the hiding strategy: the less is the information loss, the better is the data quality. This measure largely depends on the specific class of PPDM algorithms. Always a PPDM algorithm has to maintain high accuracy to reduce information loss (Aggarwal et al. 2008).

B.Completeness and Consistency

Completeness evaluates the degree of missed data in the sanitized database. Incomplete data has a significant impact on data mining results and impairs the data mining algorithms from providing an accurate representation of the underlying data. Consistency is related to the semantic constraints holding on the data and it measures how many of these constraints are still satisfied after the sanitization (Kantarcioglu et. al 2007).

C.Scalability

It is another important aspect to assess the performance of a PPDM algorithm. In particular, scalability describes the efficiency trends when data sizes increase. Such parameter concerns the increase of both performance and storage requirements as well as the costs of the communications required by a data mining technique with the increase of data size (Bertino et. al 2008).

ISSN 2278 - 1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 1, MARCH 2012

D. Data quality

It is an important aspect of PPDM. High quality data that has been prepared specifically for data mining tasks will result in useful data mining models and output. Alternatively, low quality data has a significant negative impact on the utility of data mining results (Bettini et. al 2009).

E.Security

It is the degree of protection against danger, damage, loss, and crime. There are two main approaches regarding how to deal with the problems of privacy that arise today. The first is a legal and policy approach whereby organizations are limited in how they store and use data based on privacy law and public policy. It typically works by evaluating scenarios and deciding if the privacy breach caused by using the data in a given way is justified or not. The second approach is technological, and provides enforced privacy guarantees through cryptographic means. This approach has the capability of enabling the data to be used while preventing privacy breaches (Nan Zhang et. al 2009).

VII.CONCLUSION

In this paper, we presented different PPDM techniques , requirements and issues and reiterate naïve privacy preserving methods to distribute ones and the methods for handling horizontally and vertically partitioned data. While all the purposed methods are only approximate to our goal of privacy preservation, we need to further perfect those approaches or develop some efficient methods.

REFERENCES

1. Agrawal D., and Aggarwal C.C (2007), 'On the Design and Quantification of Privacy Preserving Data Mining Algorithms', Proceedings of the 20th ACM Symposium on Principles of Database Systems, pp. 247-255.
2. Agrawal, R., and Srikant (2007), 'Privacy Preserving Data Mining', Proceedings of the 19th ACM International Conference on Knowledge Discovery and Data Mining, Canada, pp. 439-450.
3. Benjamin C. Fung M. and Ke Wang (2010), 'Privacy-Preserving Data Publishing: A Survey of Recent Developments', ACM Computing Surveys, Vol. 42, No. 4, pp.322-435.
4. Bertino E., Nai Fovino and Parasiliti Provenza (2005), 'A Framework for Evaluating Privacy Preserving Data Mining Algorithms', Journal of Data Mining and Knowledge Discovery, pp. 78-87.
5. Bikramjit Saikia and Debkumar Bhowmik (2009), 'Study of Association Rule Mining and different hiding Techniques', PhD thesis, Department of computer Science Engineering, National Institute of Technology, pp.55-63.
6. Bo Peng and Xingyu Geng (2010), 'Combined Data Distortion Strategies for Privacy-Preserving Data Mining', Proceedings of the 3rd International Conference on Advanced Computer Theory and Engineering, pp. 241-253.3
7. Chris Clifton and Murat Kantarcioglu and Jaideep Vaidya (2002), 'Defining Privacy for Data Mining', Proceedings of the National Science Foundation Workshop on Next Generation Data Mining, pp.274-281.
8. Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya (2004), 'Tools for privacy preserving data mining', Explorations Journal, volume 4, pages 28-34.
9. Clifton C., Kantarcioglu, M. and Vaidya, J. (2008), 'Defining Privacy For Data Mining', Proceedings of the National Science Foundation Workshop on Next Generation Data Mining pp. 126-133.
10. Elisa bertino and Igor fovino (2005), 'A Framework for Evaluating Privacy Preserving Data Mining Algorithms', International Journal of Data Mining and Knowledge Discovery, pp.121-154.
11. Evfimevski A., Gehrke J., and Srikant R.(2007), 'Limiting Privacy Breaches in Privacy Preserving Data Mining', Proceedings of the International Conference on Databases, pp. 171-182.
12. Evfimievski A., Srikant R., Agrawal R. and Gehrke J. (2007), 'Privacy Preserving Mining of Association Rules', Proceedings of the 8th ACM International Conference on Knowledge, Discovery and Data Mining, pp.217-228.
13. Evfimievski M. (2008), 'Randomization in Privacy Preserving Data Mining', Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining, pp.43-48.
14. Felty A. P. and Matwin S. (2007), 'Privacy-Oriented Data Mining by Proof Checking', Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery, pp.138-149.
15. Gayatri Nayak and Swagatika Devi (2011), 'A Survey On Privacy Preserving Data Mining: Approaches And Techniques', International Journal of Engineering Science and Technology, pp.2127-2133.
16. Igor Fovino and Marcelo Masera (2008), 'Privacy Preserving Data Mining : A Data Quality Approach', JRC Scientific and Technical Reports, Vol. 2, pp. 28-37.
17. Islam M. and Brankovic L. (2010), 'Noise Addition for Protecting Privacy in Data Mining', Proceedings of the 6th Engineering Conference on Mathematics and Applications, pp.207-219.
18. Islan Md. Z. and Brankovic L (2008), 'A Framework for Privacy Preserving Data Mining', Proceedings of the Australasian Workshop on Data Mining and Web Intelligence, pp. 163-168.
19. Lindell Y. and Pinkas B. (2004),'Privacy Preserving Data Mining', Journal of Cryptology, vol. 15, pp. 177-206, International Journal of Database and Data Mining, pp. 178-191.
20. Liu L., Kantarcioglu M. and B. Thuraisingham (2008), 'The applicability of the perturbation based privacy preserving data mining for real-world data', Data and Knowledge Engineering, Vol. 65, pp. 5-21.
21. Liu L., Wang J. and Zhang J. (2008), 'Wavelet based data perturbation for simultaneous privacy preserving and statistics preserving', Proceedings of IEEE International Conference on Data Mining Workshop, pp. 316-319.