International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 20 Issue 3 – MARCH 2016.

INVESTIGATING PRIVACY PRESERVATION ISSUES IN DATA MINING

Dr. S. Manimekalai[#1], M Ranjani[*2]

[#1]*Head, Department of Information Technology, Thanthai Hans Roever College, Perambalur*
[*2]*Research Scholar in Computer Science, Thanthai Hans Roever College, Perambalur*

***Abstract*--Privacy Preserving Data Mining Systems is to propose local data mining and global data mining. It attempts to benefit of extracting useful information from large volumes of data. Privacy-preserving data mining usually has multiple steps that translate to a three-tiered architecture. Online data collection systems are an example of new applications that threaten individual privacy. Already companies are sharing data mining models to obtain a richer set of data about mutual customers and their buying habits as Data Providers, Data Warehouse Server and Data Mining server. Our goal in investigating privacy preservation issues was to take a systemic view of architectural requirements and design principles and explore possible solutions that would lead to guidelines for building practical privacy preserving Central to the strategy are three protocols that govern privacy disclosure among entities as Data collection protocol, Inference Control Protocol and Information sharing Protocol.**

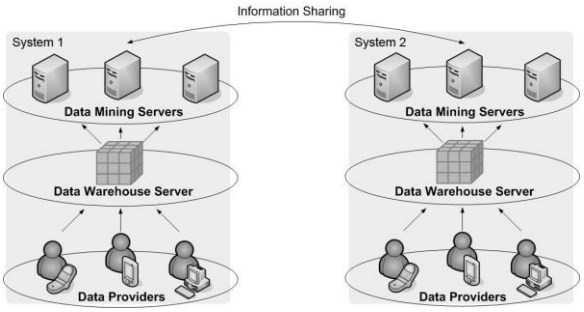
Keywords: Privacy Preserving, Inference Control Protocol, Data Warehouse, Investigating.

I. INTRODUCTION

Data mining is the process of extracting knowledge from large amounts of data. It has been widely and successfully used for more than ten years in various domains, such as marketing, weather forecasting, medical diagnostics, anti-terror measures, etc. Nonetheless, the challenge remains to conduct data mining over private data (e.g., health information) without violating the privacy of data owners (e.g., patients). Privacy protection has become a necessary requirement in many data mining applications due to emerging privacy legislation and regulations, such as the U.S.
Health Insurance Portability and Accountability Act (HIPAA) and the European Union's Privacy Directive. This dissertation seeks to design and compare strategies for protecting privacy in data mining.

I.1 Baseline Architecture

Data mining is usually carried out in multiple steps. First, the data being mined are collected from their sources, which we refer to as data providers. In many systems, data providers are physically distributed, forming the bottom tier of the baseline architecture of data mining systems, as shown in Figure 1. Data providers are the data owners, and are expected to submit their (private) data to the data warehouse server, which forms the middle tier of the architecture. For example, in an online survey system, the survey respondents are the data providers who submit their data to the survey analyzer, which holds the data warehouse server.



In the data warehouse server, data collected from the data providers are stored in well disciplined physical structures (e.g., multi-dimensional data cube), and are aggregated and pre-computed in various forms (e.g., sum, max, min). For example, in an online survey system, an aggregated data point may be the mean age of all survey respondents. The objective of data warehouse server is to support online analytical processing (OLAP) on the data, and to facilitate data mining. The actual data mining tasks are performed by the data mining servers, which form the top tier of the baseline architecture.
When performing data mining tasks, the data mining servers are likely to use the aggregated data, which are pre-computed by the data warehouse server, rather than the rough data, which are directly collected from the data providers, in order to hasten the data mining process.
Note that the data mining servers may not have the right to access all data stored in the data warehouse. For example, in a hospital where all patients' information is stored in the data warehouse, the accounting department of the hospital (as a data mining server) is allowed to access patients' financial data, but is prohibited from accessing patients medical records per HIPAA requirements. Besides constructing data mining models on its local data warehouse server, a data mining server may also share information with data mining servers from other systems (i.e., with other data warehouses), in order to construct data mining models spanning multiple data warehouses. Since each data mining server holds the local data mining model of its own system, in the information sharing

process, each data mining server is likely to share its local data mining model, rather than the raw data stored in the data warehouse, to build globally valid data mining models across multiple systems. For example, several retail companies may share their local data mining models on customer records in order to build a global data mining model on consumer behavior. Note that the local data mining models can be private and need to be protected, especially when these models are not valid globally.

I.2 Design Principle

In order to introduce the design principle of privacy-preserving data mining systems, we need to define the term "privacy". Privacy has been a central issue from a sociological standpoint. In the context of information privacy, information is considered to be private if its owner has the *right to choose* whether or not, to what extent, and for what purpose, to disclose the information to others. In the literature on privacy preserving data mining, it is commonly (explicitly or tacitly) assumed that a data owner generally chooses not to disclose its private data unless the disclosure is necessary for the purpose of data mining. Based on this assumption, we can state the design principle of privacy-preserving data mining systems as follows.

Note that the "*minimum*" here is a qualitative measure rather than a quantitative one. Since the quantitative measure of privacy disclosure varies between different systems and/or different data owners, we use the term "minimum" in the design principle to state that all private information unnecessary (or less necessary, as determined by the sensitivity of data and the accuracy requirements of data mining results) for data mining should *not* be disclosed in a privacy-preserving data mining system.

Due to the minimum necessary rule, the privacy disclosure in data mining systems should be allowed on a "need-to-know" (i.e., necessary-for-data mining) basis. The minimum necessary rule has been defined and mandated by privacy legislation and regulations. In particular, it is considered to be the key regulation of HIPAA privacy rules.

I.3 Basic Strategy

Based on the system architecture and design principle, we now introduce the basic design strategies for privacy-preserving data mining systems. Apparently, in a data mining system, privacy disclosure can occur when private data are transmitted from one entity to another. Thus, a commonly used privacy protection measure is to enforce privacy-preserving communication protocols between different entities, such that each entity may follow the protocol and thereby prevent private information disclosure during data communication.

Specifically, three kinds of protocols are needed: *Data Collection Protocol*, *Inference Control Protocol*, *Information Sharing Protocol*,

I.4 Dissertation Organization

The rest of this dissertation is organized as follows. In the second chapter, we briefly review the related work in privacy-preserving data mining. Then, we address the design of the three protocols. We propose a new scheme on data collection protocol in Chapter III. We introduce a cardinality-based inference control protocol in Chapter IV. In Chapter V, we present the adversary models and design strategies of information sharing protocols. We address the effective integration of these three protocols in Chapter VI, and conclude with final remarks in Chapter VII.

II. RELATED WORK

There has been a growing amount of research in the area of privacy-preserving data mining. In this chapter, we briefly review related work on data collection protocol, inference control protocol, and information sharing protocol, respectively. Before that, we remark that the readers should not mistake our review in this chapter as an indicator that practical privacy-preserving data mining systems have been well developed and widely used. In fact, although there is ongoing work on the development of real privacy preserving data mining systems [34], most work reviewed in this chapter presents proposals for privacy-preserving algorithms, rather than solutions to real system building problems.

2.1 Data Collection Protocol

There are three kinds of approaches that have been proposed for data collection protocol: data exchange approach, noise insertion approach and cryptographic approach.When the data exchange approach is used, each data provider exchanges its data with another data provider before transmitting the data to the data warehouse server. As such, the data warehouse server does not know the real owners of the collected data. Nonetheless, the data collected by the data warehouse server are still able to support construction of data mining models. The data exchange approach divulges the private information of one data provider to (at least) another data provider. Thus, this approach can only be used in systems where every data provider is trustworthy. That is, no data provider has the intent to compromise the private information of another data provider. Note that in many practical systems (e.g., online survey), the data providers are untrustworthy (i.e., one data provider may intend to compromise the private data of another data provider). Apparently, the data exchange approach cannot protect the private information of data providers from being compromised (by other data providers) in these systems.

2.2 Inference Control Protocol

There are two kinds of approaches that have been proposed for inference control protocol, namely the query-oriented approach [57] and the data-oriented approach. To describe the query-oriented approach, we first need to introduce a concept called "*safe*" query set. A set of queries {Q1, Q2, …, Q*n* } is safe if a data mining server cannot infer private information from the answers to Q1, Q2, …, Q*n*. With this concept, the basic idea of query-oriented inference control approach can be easily described as follows. Upon receiving a query from a data mining server, the data warehouse server will answer the query if and only if the union set of query history (i.e., the set of all the queries already answered) and the recently received query is safe. Otherwise, the data warehouse server will reject the query. Query-oriented inference control has also been extensively used in statistical databases. The major difference between these systems and privacy-preserving data mining systems is that privacy-preserving data mining systems usually deal with larger amounts of data in a timely manner (recall that OLAP means *online* analytical processing).

Information Sharing Protocol

Most existing work on information sharing considers the privacy-preserving information sharing problem as a variation of the secure multiparty computation (SMC) problem, and use cryptographic approaches to solve the problem. Since it is difficult to achieve security against adversaries with unrestricted behavior in SMC most existing information sharing protocols make restrictions on the behavior of adversaries. There are two kinds of restrictions that are commonly employed: 1) *semi-honest* (i.e., honest-but-curious) restriction, which assumes that all adversaries properly follow the protocol, with the only exception being that the adversaries may record all intermediate computation and communication and 2) *malicious* restriction, which assumes that an adversary may deviate from the protocol, but cannot change its input.

III. METHOD

3.1 Design for Association Rule Mining

We now implement our scheme to support privacy-preserving mining of association rules. Recall that there are two basic components in our scheme: 1) the perturbation guidance component of the data warehouse server, which computes the current system privacy disclosure level k^* and the perturbation guidance V_k^* , and 2) the perturbation component of the data providers, which validates V_k^* and perturbs the private data. We will present the implementation of these components for privacy-preserving association rule mining systems after introducing notions of the private dataset.

3.2 Basic Notions

Let there be *m* data providers in the system, each of which holds a private transaction *ti* (*i*∈ [1, *m*]). Let *I* be a set of *n* items *a*1, …, *an*. Each transaction *ti* is a set of items such that *ti* ∈ *I*. The data warehouse server has no external knowledge about the private information of data providers. We represent each transaction by an *n*-dimensional binary vector *ti* such that the *j*-th bit of the vector is 1 if and only if *aj* ∈ *ti*. Correspondingly, we represent the set of all private data tuples by an $m \times n$ matrix $T = [t1; \ldots; tm]$.[3] Each transaction *ti* is represented by a row vector in *T*. We denote the transpose of *T* by *T ′*. We use ⟨*T*⟩*ij* to denote the element of *T* with indices *i* and *j*.

3.3 Association Rule Mining

We first compare the performance of our scheme with that of the randomization approach in association rule mining. We use a real dataset "BMS Web view 1" from Blue Martini Software. The dataset contains several months' click stream data from Gazelle.com, a leg-care web retailer that no longer exists. We choose this dataset because it has been extensively used (e.g., in KDD Cup 2000) to test the real-world performance of association rule mining algorithms [64]. The dataset includes 59,602 transactions and 497 items. The maximum transaction size (i.e., number of items in a transaction) is 267. There is no missing value in the dataset.

3.4 Inference Control Protocol and Information Sharing Protocol

Inference control protocol and information sharing protocol are normally transparent to each other, as the inference control protocol enables a data mining server to construct local data mining models, and the information sharing protocol enables a data mining server to share local data mining models and construct global data mining models spanning multiple systems. Nonetheless, there are certain cases where inference control protocol and information sharing protocol need to be integrated with each other. Recall that in order for the information sharing protocol to work, the participating parties (i.e., data mining servers from different systems) must have a specifically designed cryptographic algorithm for every data mining task. In cases where such a specific algorithm is unavailable, a possible alternative is for each party to allow other parties (from other systems) to directly access its local data warehouse. In this case, the privacy protection must be implemented in the inference control protocol to accommodate the requirements of information sharing. The objective of the (new) inference control protocol becomes 1) to allow local data mining servers to learn the minimum private information necessary for data mining, and 2) to prevent remote data mining servers of other systems from inferring private information stored in the data warehouse.

IV. CONCLUSION

We now conclude this dissertation with a discussion on open issues that need to be addressed in order to further improve the performance of privacy-preserving data mining techniques. Heterogeneous Privacy Requirements: The design of privacy-preserving data mining techniques depends on the specification of privacy protection levels required by the data owners. Most existing studies assume (at least partially)

homogenous privacy requirements. That is, all data owners have the same level of privacy requirement on all of their data and/or all attributes. This assumption simplifies the design and implementation of privacy-preserving data mining techniques, but cannot reflect the privacy concerns in practice. Indeed, due to multiple survey different people have diversified privacy requirements on different data and/or different attributes. The work in this dissertation (e.g., our scheme for data collection protocol) removes part of the assumption, as we allow different data providers to specify different levels of privacy protection on their data. Nonetheless, we still cannot allow a data provider to explicitly assign different privacy protection levels on different attributes. It would be challenging, but potentially beneficial, to design and implement new techniques that fully address the heterogeneous privacy requirements of data owners.

REFERENCES

[1] D. Agrawal and C. C. Aggarwal, "On the Design and Quantification of Privacy Preserving Data Mining Algorithms," *Proc. 20th ACM SIGMOD-SIGACTSIGART Symposium on Principles of Database Systems*, pp. 247-255, 2001.

[2] R. Agrawal, D. Asonov, and R. Srikant, "Enabling Sovereign Information Sharing Using Web Services," *Proc. 23rd ACM SIGMOD International Conference on Management of Data*, pp. 873-877, 2004.

[3] R. Agrawal, A. Evfimievski, and R. Srikant, "Information Sharing Across Private Databases," *Proc. 22nd ACM SIGMOD International Conference on Management of Data*, pp. 86-97, 2003.

[4] R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining," *Proc. 19th ACM SIGMOD International Conference on Management of Data*, pp. 439-450, 2000.

[5] R. Agrawal, R. Srikant, and D. Thomas, "Privacy Preserving OLAP," *Proc. 25th ACM SIGMOD International Conference on Management of Data*, pp. 251 - 262, 2005.

[6] S. Agrawal and J. Haritsa, "A Framework for High-Accuracy Privacy-Preserving Mining," *Proc. 21st International Conference on Data Engineering*, pp. 193-204, 2005.

[7] S. Agrawal, V. Krishnan, and J. Haritsa, "On Addressing Efficiency Concerns in Privacy-Preserving Mining," *Proc. Ninth International Conference on Database Systems for Advanced Applications*, pp. 113-124, 2004.

[8] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. New York, NY: Addison-Wesley, 1999.

[9] R. J. Bayardo and R. Agrawal, "Data Privacy through Optimal *k*-Anonymization," *Proc. 21st International Conference on Data Engineering*, pp.217-228, 2005.

[10] C. Blake and C. Merz, *UCI repository of machine learning databases*. Irvine, CA: University of California, Department of Information and Computer Science, 1998.