Chapter 2.3
Privacy-Preserving Data Mining: Development and Directions

Bhavani Thuraisingham
The MITRE Corporation, USA

ABSTRACT

This article first describes the privacy concerns that arise due to data mining, especially for national security applications. Then we discuss privacy-preserving data mining. In particular, we view the privacy problem as a form of inference problem and introduce the notion of privacy constraints. We also describe an approach for privacy constraint processing and discuss its relationship to privacy-preserving data mining. Then we give an overview of the developments on privacy-preserving data mining that attempt to maintain privacy and at the same time extract useful information from data mining. Finally, some directions for future research on privacy as related to data mining are given.

INTRODUCTION

There has been much interest recently on applying data mining for counter-terrorism applications (see Thuraisingham, 2003a, 2003b). For example, data mining can be used to detect unusual patterns, terrorist activities and fraudulent behavior. While all of these applications of data mining can benefit humans and save lives, there is also a negative side to this technology, since it could be a threat to the privacy of individuals. This is because data mining tools are available on the Web or otherwise, and even naive users can apply these tools to extract information from the data stored in various databases and files, and consequently violate the privacy of individuals. As we have stressed in other papers (see Thuraisingham, 2003a), to carry out

effective data mining and extract useful information for counter-terrorism and national security, we need to gather all kinds of information about individuals. However, this information could be a threat to individuals' privacy and civil liberties (Thuraisingham, 2002).

Privacy is getting more attention partly because of counter-terrorism and national security. Recently we have heard a lot about national security in the media. This is mainly because people are now realizing that to handle terrorism, the government may need to collect information about individuals. This is causing a major concern with various civil liberties unions. The challenge is to carry out data mining and yet maintain privacy. This topic is known as privacy-preserving data mining.

This paper discusses developments and directions for privacy-preserving data mining, also sometimes called privacy sensitive data mining or privacy enhanced data mining. We discuss the privacy problem, provide an overview of the developments in privacy-preserving data mining and then discuss some of our research on viewing the privacy problem as an inference problem. In the next section, we first provide an overview of the privacy problem and discuss the connection between the privacy problem and the inference problem. Our research on developing techniques for ensuring privacy follows. This approach is called privacy constraint processing. We also show the connection between privacy-constraint processing and privacy-preserving data mining. Developments in privacy-preserving data mining will be discussed afterwards, along with directions for privacy research.

PRIVACY, DATA MINING AND THE INFERENCE PROBLEM

With the World Wide Web, there is now an abundance of information about individuals that one can obtain within seconds. This information could be obtained through mining or just from information retrieval. Data mining is the process of users posing queries and extracting information previously unknown using machine learning and other reasoning techniques (see Thuraisingham, 1998). Now, data mining is an important technology for many applications. However data mining also causes privacy concerns, as users can now put pieces of information together and extract information that is sensitive or private. Therefore, one needs to enforce controls on databases and data mining tools. That is, while data mining is an important tool for many applications, we do not want the information extracted to be used in an incorrect manner. For example, based on information about a person, an insurance company could deny insurance or a loan agency could deny loans. In many cases these denials may not be legitimate. Therefore, information providers have to be very careful in what they release. Also, data mining researchers have to ensure that privacy aspects are addressed.

We are beginning to realize that many of the techniques that were developed for the past two decades or so on the inference problem can now be used to handle privacy. One of the challenges to securing databases is the inference problem (Air Force Science Board, 1983). Inference is the process of users posing queries and deducing unauthorized information from the legitimate responses that they receive. This problem has been discussed quite a lot over the past two decades (Thuraisingham, 1987; Morgenstern, 1987; Hinke, 1988). However, data mining makes this problem worse. Users now have sophisticated tools that they can use to get data and deduce patterns that could be sensitive. Without these data mining tools, users would have to be fairly sophisticated in their reasoning to be able to deduce information from posing queries to the databases. That is, data mining tools make the inference problem quite dangerous (Clifton & Marks, 1996). While the inference problem mainly deals with secrecy and confidentiality, we are beginning to see many