Homeland Security Data Mining and Link Analysis

Bhavani Thuraisingham
The MITRE Corporation, USA

INTRODUCTION

Data mining is the process of posing queries to large quantities of data and extracting information often previously unknown using mathematical, statistical, and machine-learning techniques. Data mining has many applications in a number of areas, including marketing and sales, medicine, law, manufacturing, and, more recently, homeland security. Using data mining, one can uncover hidden dependencies between terrorist groups as well as possibly predict terrorist events based on past experience. One particular data-mining technique that is being investigated a great deal for homeland security is link analysis, where links are drawn between various nodes, possibly detecting some hidden links.

This article provides an overview of the various developments in data-mining applications in homeland security. The organization of this article is as follows. First, we provide some background on data mining and the various threats. Then, we discuss the applications of data mining and link analysis for homeland security. Privacy considerations are discussed next as part of future trends. The article is then concluded.

BACKGROUND

We provide background information on both data mining and security threats.

Data Mining

Data mining is the process of posing various queries and extracting useful information, patterns, and trends often previously unknown from large quantities of data possibly stored in databases. Essentially, for many organizations, the goals of data mining include improving marketing capabilities, detecting abnormal patterns, and predicting the future, based on past experiences and current trends. There is clearly a need for this technology. There are large amounts of current and historical data being stored. Therefore, as databases become larger, it becomes increasingly difficult to support decision making. In addition, the data could be from multiple sources and multiple domains. There is a clear need to analyze the data to support planning and other functions of an enterprise.

Some of the data-mining techniques include those based on statistical reasoning techniques, inductive logic programming, machine learning, fuzzy sets, and neural networks, among others. The data-mining problems include classification (finding rules to partition data into groups), association (finding rules to make associations between data), and sequencing (finding rules to order data). Essentially, one arrives at some hypotheses, which is the information extracted from examples and patterns observed. These patterns are observed from posing a series of queries; each query may depend on the responses obtained to the previous queries posed.

Data mining is an integration of multiple technologies. These include data management, such as database management, data warehousing, statistics, machine learning, decision support, and others, such as visualization and parallel computing. There is a series of steps involved in data mining. These include getting the data organized for mining, determining the desired outcomes to mining, selecting tools for mining, carrying out the mining, pruning the results so that only the useful ones are considered further, taking actions from the mining, and evaluating the actions to determine benefits. There are various types of data mining. By this we do not mean the actual techniques used to mine the data, but what the outcomes will be. These outcomes also have been referred to as data-mining tasks. These include clustering, classification anomaly detection, and forming associations.

While several developments have been made, there also are many challenges. For example, due to the large volumes of data, how can the algorithms determine which technique to select and what type of data mining to do? Furthermore, the data may be incomplete and/or inaccurate. At times, there may be redundant information, and at times, there may not be sufficient information. It is also desirable to have data-mining tools that can switch to multiple techniques and support multiple outcomes. Some of the current trends in data mining include mining Web data, mining distributed and heterogeneous databases, and privacy-preserving data mining, where one ensures that one can get useful results from mining and at the same time maintain the privacy of

individuals (Berry & Linoff; Han & Kamber, 2000; Thuraisingham, 1998).

Security Threats

Security threats have been grouped into many categories (Thuraisingham, 2003). These include information-related threats, where information technologies are used to sabotage critical infrastructures, and non-information-related threats, such as bombing buildings. Threats also may be real-time threats and non-real-time threats. Real-time threats are threats where attacks have timing constraints associated with them, such as "building X will be attacked within three days." Non-real-time threats are those threats that do not have timing constraints associated with them. Note that non-real-time threats could become real-time threats over time.

Threats also include bioterrorism, where biological and possibly chemical weapons are used to attack, and cyberterrorism, where computers and networks are attacked. Bioterrorism could cost millions of lives, and cyberterrorism, such as attacks on banking systems, could cost millions of dollars. Some details on the threats and countermeasures are discussed in various texts (Bolz, 2001). The challenge is to come up with techniques to handle such threats. In this article, we discuss data-mining techniques for security applications.

MAIN THRUST

First, we will discuss data mining for homeland security. Then, we will focus on a specific data-mining technique called *link analysis* for homeland security. An aspect of homeland security is cyber security. Therefore, we also will discuss data mining for cyber security.

Applications of Data Mining for Homeland Security

Data-mining techniques are being examined extensively for homeland security applications. The idea is to gather information about various groups of people and study their activities and determine if they are potential terrorists. As we have stated earlier, data-mining outcomes include making associations, linking analyses, forming clusters, classification, and anomaly detection. The techniques that result in these outcomes are techniques based on neural networks, decisions trees, market-basket analysis techniques, inductive logic programming, rough sets, link analysis based on graph theory, and nearest-neighbor techniques. The methods used for data mining include top-down reasoning, where we start with a hypothesis and then determine whether the hypothesis is true, or bottom-up reasoning, where we start with examples and then come up with a hypothesis (Thuraisingham, 1998). In the following, we will examine how data-mining techniques may be applied for homeland security applications. Later, we will examine a particular data-mining technique called *link analysis* (Thuraisingham, 2003).

Data-mining techniques include techniques for making associations, clustering, anomaly detection, prediction, estimation, classification, and summarization. Essentially, these are the techniques used to obtain the various data-mining outcomes. We will examine a few of these techniques and show how they can be applied to homeland security. First, consider association rule mining techniques. These techniques produce results, such as John and James travel together or Jane and Mary travel to England six times a year and to France three times a year. Essentially, they form associations between people, events, and entities. Such associations also can be used to form connections between different terrorist groups. For example, members from Group A and Group B have no associations, but Groups A and B have associations with Group C. Does this mean that there is an indirect association between A and C?

Next, let us consider clustering techniques. Clusters essentially partition the population based on a characteristic such as spending patterns. For example, those living in the Manhattan region form a cluster, as they spend over $3,000 on rent. Those living in the Bronx from another cluster, as they spend around $2,000 on rent. Similarly, clusters can be formed based on terrorist activities. For example, those living in region X bomb buildings, and those living in region Y bomb planes.

Finally, we will consider anomaly detection techniques. A good example here is learning to fly an airplane without wanting to learn to take off or land. The general pattern is that people want to get a complete training course in flying. However, there are now some individuals who want to learn to fly but do not care about take off or landing. This is an anomaly. Another example is John always goes to the grocery store on Saturdays. But on Saturday, October 26, 2002, he went to a firearms store and bought a rifle. This is an anomaly and may need some further analysis as to why he is going to a firearms store when he has never done so before. Some details on data mining for security applications have been reported recently (Chen, 2003).

Applications of Link Analysis

Link analysis is being examined extensively for applications in homeland security. For example, how do we