Chapter XLI
Data Mining

Mark Last
Ben-Gurion University of the Negev, Israel

ABSTRACT

Data mining is a growing collection of computational techniques for automatic analysis of structured, semi-structured, and unstructured data with the purpose of identifying important trends and previously unknown behavioral patterns. Data mining is widely recognized as the most important and central technology for homeland security in general and for cyber warfare in particular. This chapter covers the following relevant areas of data mining:

- *Web mining is the application of data mining techniques to web-based data. While Web usage mining is already used by many intrusion detection systems, Web content mining can lead to automated identification of terrorist-related content on the Web.*
- *Web information agents are responsible for filtering and organizing unrelated and scattered data in large amounts of web documents. Agents represent a key technology to cyber warfare due to their capability to monitor multiple diverse locations, communicate their findings asynchronously, collaborate with each other, and profile possible threats.*
- *Anomaly detection and activity monitoring. Real-time monitoring of continuous data streams can lead to timely identification of abnormal, potentially criminal activities. Anomalous behavior can be automatically detected by a variety of data mining methods.*

INTRODUCTION

Data mining (DM) is a rapidly growing collection of computational techniques for automatic analysis of structured, semi-structured, and unstructured data with the purpose of identifying various kinds of previously unknown behavioral patterns. According to Mena (2004), data mining is widely recognized as the most important and central technology for homeland security in general and for cyber warfare in particular. This relatively new field emerged in the beginning of the 1990s as a combination of methods and algorithms from statistics, pattern recognition, and machine learning. The difference between data mining and knowledge discovery in databases (KDD) is defined by as follows: data mining refers to the application of pattern extraction algorithms to data, while KDD is the overall process of "identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad, Piatetsky-Shapiro & Smyth, 1996, p. 6). The complete KDD process includes such stages as data selection; data cleaning and pre-processing; data reduction and transformation; choosing data mining tasks, methods and tools; data mining (searching for patterns of ultimate interest); interpretation of data mining results; and action upon discovered knowledge.

BACKGROUND

Tens of computational techniques related to various data mining tasks emerged over the last 15 years. Selected examples of some common data mining tasks and algorithms will be briefly described.

Association rules: Association rule mining is aimed at finding interesting association or correlation relationships among a large set of data items (Han & Kamber, 2001). The extracted patterns (association rules) usually have the form "if event *X* occurs, then event *Y* is likely." Events *X* and *Y* may represent items bought in a purchase transaction, documents viewed in a user session, medical symptoms of a given patient, and many other phenomena recorded in a database over time. Extracted rules are evaluated by two main parameters: *support*, which is the probability that a transaction contains both *X* and *Y* and *confidence*, which is the conditional probability that a transaction having *X* also contains *Y*. Scalable algorithms, such as *Apriori* (Srikant & Agrawal, 1996), have been developed for mining association rules in large databases containing millions of multi-item transactions.

Cluster analysis: A *cluster* is a collection of data objects (e.g., Web documents) that are similar to each other within the same cluster, while being dissimilar to the objects in any other cluster (Han & Kamber, 2001). One of the most important goals of cluster analysis is to discover *hidden patterns*, which characterize groups of seemingly unrelated objects (transactions, individuals, documents, etc.). Clustering of "normal walks of life" can also serve as a basis for the task of *anomaly detection*: an outlier, which does not belong to any normal cluster, may be an indication of abnormal, potentially malicious behavior (Last & Kandel, 2005, chap. 4 & 6). A survey of leading clustering methods is presented in *Data Clustering: A Reveiw* (Jain, Murty, & Flynn, 1999).

Predictive modeling: The task of *predictive modeling* is to predict (anticipate) future outcomes of some complex, hardly understandable processes based on automated analysis of historic data. Predicting future behaviors (especially attacks) of terrorist and other malicious groups is an example of such task. Han and Kamber (2001) refer to prediction of continuous values as *prediction*, while prediction of nominal class labels (e.g., terrorist vs. non-terrorist documents) is regarded by them as *classification*. Common classification models include ANN—Artificial Neural Networks (Mitchell, 1997), decision trees (Quinlan, 1993), Bayesian networks (Mitchell, 1997), IFN—Info-Fuzzy Networks (Last & Maimon, 2004), and so forth.

Visual data mining: *Visual data mining* is the process of discovering implicit but useful knowledge from large data sets using visualization techniques. Since "a