



JASON ANDRESS

Data Mining as a Tool for Security

Difficulty



Given the current state of heightened security across the globe today, the ability to sift through data and search for key information and the occurrence of particular patterns is highly desirable.

This capability, known as data mining, can be used to pinpoint anything from seasonal grocery purchasing habits of individuals to the pattern of international telephone calls that might presage an act of terrorism.

Data mining, simply, is a process that allows large volumes of data to be searched for patterns and relationships in or among sets of data.

Goals of Data Mining

Data mining is performed with the return of particular information in mind. In a general sense, the ends to the means of data mining are broken into four main categories:

Prediction

Data mining can be used in an attempt to predict how certain variables in the data set will behave in the future. This type of analysis is often used in intrusion detection systems. Being able to predict what network traffic will look like on a particular day or at a particular time of day allows security personnel to focus their efforts on the specific areas where unusual activity is noted.

Discovery

Patterns within mined data can be used to identify the existence of a given item, event, or activity. In a general sense, password authentication fits within this category. Password authentication determines whether a user is actually a specific user by comparing the

attributes that the user presents against the stored attributes held in a database.

Classification

Mined data can be partitioned in order to identify classes or categories based on combinations of parameters. Such methods are often used to segment network traffic into 'good' and 'bad' based on the contents of a particular packet, as is done in simple packet filtering, or on the contents of a packet in context with other traffic, in order to detect more complex behavior such as IP fragmentation attacks.

Optimization

Data mining can be used in an effort to optimize the use of resources in a variety of projects. An excellent example of data mining used for this purpose is that of estimating software quality. The task of predicting software faults can be truly daunting, but with data mining techniques and information gathered from a variety of software metrics gathering methods, such faults can be accurately predicted [1].

Types of Information Returned by Data Mining

Data mining can be used as a sort of inductive reasoning, meaning that it can be used to discover previously unknown patterns within data. Four types of data can be returned by data mining:

WHAT YOU WILL LEARN...

Basic security terminology

WHAT SHOULD YOU KNOW...

The basics of data mining

Where the data comes from

How data mining is used for security

Association Rules

Association rules can be used to correlate the presence of a set of items with another range of values for another set of variables. Association rules are particularly useful when attempting to detect unauthorized activity based on log files. While some events, like the update of a critical system file, may be expected when applying a patch or an update, the same file being updated in the absence of maintenance activity may be an indicator of an attack.

Classification Hierarchies

Classification hierarchies work from existing sets of events or transactions to create a hierarchy of classes. One interesting use of this system is present in the *Computer Aided Facial Image Inference and Retrieval System* (CAFIIR) [2]. CAFIIR is a tool that assists witnesses in identifying suspects from photographs. A classification hierarchy used to allow the photographs to be ordered by facial features in order more easily identify individuals.

Sequential Patterns

Sequential patterns are patterns where a sequence of events or actions typically happens in a specific order. A motorist with a flat tire will likely follow a sequence similar to the following:

- Pull over to the side of the road
- Exit the vehicle
- Examine the tire
- Open the trunk
- Retrieve the spare tire

This sequence of events is unlikely to occur in a different order. When such patterns are discovered, accurate predictions can generally be made regarding future occurrences from any point in the sequence. Sequential pattern analysis is used when examining communications traffic for signs of impending terrorist attacks and has been found to be a reliable indicator of such activity, even clearly indicating past attacks where communications logs were available for analysis afterward.

Periodicity

Time series data examines the change of data values over a period of time. Periodicity is used to help predict the behavior of time series data. A very common use for this sort of data is found in tracking the behavior of malware. Data mining, based on time series data, can be used to assist in predicting the occurrence and behavior of malware, and to model it as it spreads.

Clustering

A population of events or items can be partitioned into sets of similar events. One example is found in a subset of clustering, known as document clustering. Document clustering, used heavily in information retrieval, text mining, databases, and many other fields, involves grouping similar or related documents together. This allows documents to be more easily searched by related topics and greatly speeds the effort of searching for particular information.

Where Does the Data Come From?

The supply of fodder for data mining can come from a variety of sources, as nearly every move made in the modern world has an associated record of some sort. Even when deliberately attempting to minimize the production of such records, individuals are still registered in multiple systems daily unless they have dropped completely off of the grid. Taking privacy measures to this extent is untenable to most people.

Purchase Records

Purchase records are a very rich source of data for data mining purposes. Making a very data-rich situation even richer is the trend in the last few years toward *loyalty cards* or *membership cards*. These cards, in place in many large retail chains and grocery stores, give discounts to customers while allowing the retailer to track the customers purchase history at an even more granular level. This data can be used to create a profile on customers over a period of time, becoming more accurate with each purchase [3].

Travel Records

When traveling by car, it is possible to track the route by purchase patterns. If only cash is used for gas purchases and a cell phone is not used then an individual can travel freely without being tracked. When traveling by air, however, data is always generated regarding time of departure, destination, and length of stay. This information was requested from the airlines by the government recently in an attempt to track terrorists.

Insurance Records

Insurance companies make a business out of risk management. For these companies to be profitable they must take in more money in the form of premiums than they pay out in claims. Insurance companies base premium pricing on the level of risk and that risk can be computed more accurately if more information is known about the potential customer. Gerver and Barrett point out that *Plan sponsors can reduce or avoid future health care costs by at least 5% or 10% annually through the use of evidence-based data mining technology* [4]. Customers are required to fill out applications for insurance and in doing so provide obvious data such as name, address and so on. This information can be used with data mining techniques to gather other information about a potential customer such as the sports that they engage in or medical ailments they might have.

Communications Records

Communications are another very rich source of information to be mined. The *US National Security Agency* (NSA) began collecting and mining phone call records since shortly after the September 11th terrorist attacks on the world trade center. These records are examined for particular patterns of calls coming into and going out of the country, a possible indicator of terrorist activity [5].

A bill recently introduced to the US House of Representatives would require Internet Service Providers (ISPs) to keep records of user's internet usage, including web browsing, *Instant Messenger* (IM) conversations, and email, potentially

indefinitely [6]. Such a massive store of data would enable data mining on a scale that would make current efforts pale in comparison.

Most web site visits are tracked. Through the use of cookies, it is possible to track individuals across multiple sites. Advertising companies such as DoubleClick (now owned by Google) make a business of tracking web site visits by individuals.

Credit History

In the 1960s, automated credit scoring came into use. This allowed the consumers credit data to be mined in order to create an overall 'score' of the customer's credit worthiness.

One of the most commonly known sources of personal information in the United States is the data housed by the credit bureaus. The main credit bureaus in the US are Equifax, Experian, and TransUnion. The credit bureaus keep data on all of our credit activities. This includes home mortgages, car and student loans and all of our credit card payment activity. Credit reports on individuals are trivially easy to obtain and can be purchased on the internet for a small fee.

Medical Records

A patient's medical information can be legally mined by insurance companies, doctors, hospitals, and other related industries. This information, containing everything from clinic visits to information on surgeries, provides a vast wealth of information. This is one of the few areas in most countries where the patient's private data is offered some sort of legal protection.

Government Records

In the United States, the FBI has been building a national DNA database since 1990. According to data on the FBI's web site, the database contained 7,261,604 profiles as of September 2009 [7]. Under the DNA Identification Act of 1994, the FBI was authorized to create a national DNA database for law enforcement purposes. Local and state police contribute data from criminals to this database. Its purpose is to enable DNA evidence at

crime scenes to be matched quickly to a person. At the present time only convicted criminals are forced to provide DNA samples. There is no law that forces suspects to contribute their DNA but such a law is possibly coming. If passed, anyone taken into custody might be required to provide a DNA sample.

In the United Kingdom, the UK *National Criminal Intelligence Database* (NDNAD) has been constructed, against which the FBI's database pales in comparison. The UK collects DNA from samples at crime scenes, anyone detained at a police station (charged or otherwise), anyone participating in a *recordable offense*, anyone convicted of certain crimes before the establishment of the database, and the deceased. DNA samples can legally be collected with or without the consent of the individual, by force if necessary [8]. As of March of 2009 the NDNAD contained 5,208,988 profiles [9], with the UK at a population nearly 1/5 that of the United States.

Data Mining for Security

With such a rich set of data to work from, data mining is frequently performed in the name of security. These uses range from security efforts regarding individuals, to the internal functions of network security devices.

Casinos

Casinos make money based on computed odds. If a customer can find a way to tip the odds in their favor they stand to make substantial gains. Jeff Jonas founded *Systems Research & Development* (SRD) to help businesses solve difficult problems using data mining techniques. Quite a bit of Jonas' early work was spent helping hospitals identify patients who were trying to avoid payment by slightly changing the name under which they registered [10].

Casinos use SRD's products to match lists of people taken from multiple sources against anyone who applies for a job or is currently or was previously employed by the casino. By using Jonas' matching algorithms, the casinos can avoid employing people who might have criminal or suspicious backgrounds.

Such systems are also used to track cheaters and card counters and share this information between different establishments.

Government

The government of the United States is investing heavily in data mining technology as part of its war on terror and for various other purposes. In 2004, the Government Accounting Office identified 199 separate government data mining operations [11]. A significant number of these operations involved the use of personal information.

More recently the US government has employed data mining techniques to detect fraudulent claims after hurricanes Katrina and Rita [12]. Some of the data mining work done to detect fraudulent claims is similar to the work that Casino owners are doing in order to avoid employing criminals and other undesirables. In the case of the fraudulent claims, the government was particularly interested in detecting multiple claims.

Possibly the highest visibility government data mining operation has been its attempt to determine if an airline passenger poses a potential threat. The *Computer-Assisted Passenger Prescreening System II* (CAPPS II) used data mining techniques applied to large amounts of airline passenger information (obtained by the government's contractor with the airlines) to attempt to match passengers to wanted lists. This particular use of data mining met with considerable public resistance as consumers discovered that their personal information had been given to the government for testing of this system, and was dismantled by president Bush. CAPPS II is now slated to be replaced by the Secure Flight system in 2010, a system that shares a great many similarities, both positive and negative.

Background Checks

Despite the detail an applicant might provide in a resume, a potential employer would be remiss if they did not investigate candidates more thoroughly. By using data mining techniques, employers can find out a lot about a potential employee. The most rudimentary form of data mining

is to simply type the person's name into Google and scan through the results. More sophisticated techniques can provide a wealth of information that may help make a decision on the suitability of the candidate, and an entire industry exists to serve such needs.

Network Security Appliances and Anti-Malware products

As mentioned briefly earlier, data mining provides a valuable tool for use in host and network security applications. Various techniques can be utilized in intrusion detection systems, firewalls, and even simple routers to ensure that malformed or malicious traffic is quickly filtered from the network.

When examining intrusion detection systems and anti-malware products from a data mining standpoint, both are similar efforts and use similar techniques. In such systems, data mining efforts are generally categorized as either misuse detection or as anomaly detection.

Misuse detection searches for patterns of attack based on an existing signature of the attack. As with most signature based systems, this type of detection is very limited, as it is not capable of detecting attacks for which it does not have a signature.

Anomaly detection works from a baseline of normal behavior and looks for significant deviation from the baseline. In theory, anything deviating by a certain amount from the baseline should be considered an attack, but this method is prone to false positives.

In many products, misuse detection and anomaly detection are used in combination to provide the most coverage.

Software Security

With the broad acceptance of open source software and the reuse of open source code, it becomes important to know what the code is actually doing. When a vendor ships a product based around a large body of code that it did not create, such as an open source database or web server, the task of validating that the code is not performing any malicious activity can be daunting, at best. Data mining techniques

WWW.CYBER-RECON.COM

Exceptional Computer Security Training
CompTIA Security+ Training in Stafford, Virginia, USA
Stafford, Virginia Class May 30 - June 4, 2010

Class Limited to 12 students

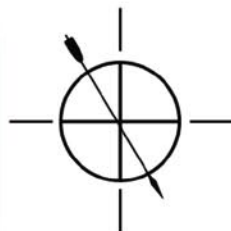
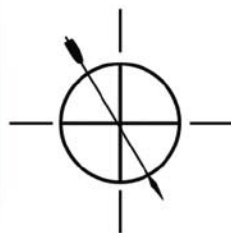
All Inclusive Training — Materials Yours To Keep After Training

• ASUS Netbook • Test Voucher • Books and Training Material • Catered Meals

Online Mentored Training Starting April 2, 2010

information@cyber-recon.com

(571) 255-2771



can be used to evaluate both the code itself, in an attempt to locate programming techniques that indicate poor or lacking security, or can be used to look for actual malicious segments of code. Several commercial tools exist that can perform these types of functions.

Privacy Issues

A great deal of data mining is based on databases of personal information such as purchases made at stores, travel plans, hobbies and medical records.

Who should have access to a person's personal information? How much personal information should they have access to? The medical and legal professions deal with client confidentiality all the time because doctors and lawyers are privy to personal information that they need to know in order to perform a service on the behalf of the patient. In the modern world, much of this information is held in digital form allowing it to be easily read, edited, copied and sent to third parties. If such data is held in a shared database, is it possible to ensure that someone other than an authorized doctor or lawyer does not have a way to read it?

Along with the confidentiality of personal data lies its integrity. Obviously it is not

desirable to allow just anyone to read private medical records and it is certainly important to prevent anyone from being able to alter these records other than to add legitimate new material. Assuring the integrity of personal data is difficult because individuals are generally not the holder of the information. Such data is held on behalf of individuals by credit rating companies, banks, doctors, lawyers and others.

Accuracy of Data

The accuracy of personal data stored by any organization can be of considerable importance to the individual.

Probably the best known problem in the US is incorrect data in an individual's credit report caused by something as trivial as a data entry error caused by a bounced check or other problem that is attributed to the wrong person. Even though that data is incorrect, proving that it is so and getting it amended becomes the responsibility of the affected individual. Families with several living persons having the same name are also prone to inaccurate credit reporting because data gets associated with the wrong family member.

The US government's recent attempts to locate terrorists by inspecting

airline manifests and so on rely on having accurate data. If an individual is unfortunate enough to have a name or address similar to a known terrorist they might find themselves denied access to a flight. The data is inaccurate – the person is not a terrorist, yet the data available to the security officials indicates that they are.

Transcription of hand-written data to computerized form is prone to errors. Consider that many patient records created by personal doctors or those doctors working in hospitals are often written by hand. When these are transcribed to a digital form, small errors in the data could have several outcomes including:

- Being billed for the treatment of another patient
- Being issued the incorrect medication or dosage
- Undergoing the wrong procedure entirely

These types of errors in medical data can be unpleasant at the very least, if not fatal.

Conclusion

The ability of modern computer systems to efficiently sift through very large amounts of data quickly has enabled the use of data mining in a wide variety of application domains. Data mining is now used to identify potential terrorists and other criminals, project sales patterns in retail stores and predict the course of the weather. An enormous amount of data is now available regarding businesses and individuals and this volume of information is increasing at an amazingly fast pace. By using data mining to gather information from multiple disparate sources, it is possible to correlate what is apparently unrelated data and to come to significant conclusions in near real-time. Data mining currently plays an increasingly important role in security, business and our personal lives, and will continue to do so for the foreseeable future.

Jason Address

Jason Address is a tinkerer, rapscallion, and all around geek. He works for a major software company, teaches graduate and undergraduate security courses, and enjoys a good game of Scrabble. He can be reached at jason.address@gmail.com.

Bibliography

- Mertik, Matej and Lenic, Mitja and Stiglic, Gregor and Kokol, Peter, Estimating Software Quality with Advanced Data Mining Techniques, 2006, <http://portal.acm.org/citation.cfm?id=1193789#>; [1]
- Wu, Jian Kang and Narasimhalu, Arcot Desai, Identifying Faces Using Multiple Retrievals, 1994, <http://dx.doi.org/10.1109/93.311656>; [2]
- Michelle Kessler and Byron Acohido, Data miners dig a little deeper, 2006, http://www.usatoday.com/tech/news/internetprivacy/2006-07-11-data-mining_.xhtml; [3]
- Barrett, H. M. Gerver and J. S.I, Data Mining-Driver ROI: Health Care Cost Management, 2006, <http://www.ifebp.org/pdf/webexclusive/06june.pdf>; [4]
- Leslie Cauley, NSA has massive database of Americans' phone calls, 2006, http://www.usatoday.com/news/washington/2006-05-10-nsa_.xhtml; [5]
- Lamar Smith, H.R.1076 – Internet Stopping Adults Facilitating the Exploitation of Today's Youth (SAFETY) Act of 2009, 2009, <http://www.opencongress.org/bill/111-h1076/show>; [6]
- Federal Bureau of Investigation, CODIS – NDIS Statistics, 2009, <http://www.fbi.gov/hq/lab/codis/clickmap.htm>; [7]
- GenWatch UK, A Brief Legal History of the NDNAD, 2009, <http://www.genewatch.org/sub-537968>; [8]
- www.parliament.uk, House of Lords Written Answers 20 April 2009, 2009, <http://services.parliament.uk/hansard/Lords/ByDate/20090420/writtenanswers/part080.html>; [9]
- Phil Becker, Finding Identity in the Noise, 2004, <http://magazine.digitalidworld.com/Mar04/Page20.pdf>; [10]
- Committee on Governmental Affairs, U.S. Senate, Data Mining. Federal Efforts Cover a Wide Range of Uses, 2004, <http://www.gao.gov/new.items/d04548.pdf>; [11]
- United States Government Accountability Office, Expedited Assistance for Victims of Hurricanes Katrina and Rita, 2006, <http://www.gao.gov/cgi-bin/gettrpt?GAO-06-655>; [12]