WEB CRIME DETECTION USING DATA MINING TECHNIQUE

[1] Dr.B.Sunil Kumar, [2]J.Jayalakshmi, [3]V.Muniraju Naidu

[1]Professor, [2] Sr.Assistant Professors, [3]Associate Professor
Dept. of CSE,
Narayana Engineering College, Nellore, A.P, India

Abstract- Globally the internet is been accessed by enormous people within their restricted domains. When the client and server exchange messages among each other, there is an activity that can be observed in log files. Log files give a detailed description of the activities that occur in a network that shows the IP address, login and logout durations, the user's behavior etc. There are several types of attacks occurring from the internet. Our focus of research in this paper is Denial of Service (DoS) attacks with the help of pattern recognition techniques in data mining. Through which the Denial of Service attack is identified. Denial of service is a very dangerous attack that jeopardizes the IT resources of an organization by overloading with imitation messages or multiple requests from unauthorized users.

Keywords- Cyber Crime, Web Crime Mining, Crime Data Mining Techniques, Forensics Analysis, Web Mining .

1. INTRODUCTION: Crime is a billion-dollar business and it is increasing every year. The PwC global economic crime survey of 2009 suggests that close to 30 percent of companies worldwide have reported being victims of crime in the past year.

Crime involves one or more persons who intentionally act secretly to deprive another of something of value, for their own benefit. Fraud is as old as humanity itself and can take an unlimited variety of different forms. However, in recent years, the development of new technologies has also provided further ways in which criminals may commit crime. In addition to that, business reengineering, reorganization or downsizing may weaken or eliminate control, while new information systems may present additional opportunities to commit crime.Crime analysis can occur at various levels, including tactical, operational, and strategic. Crime analysts study crime reports, arrests reports, and police calls for service to identify emerging patterns, series, and trends as quickly as possible. They analyze these phenomena for all relevant factors, sometimes predict or forecast future occurrences, and issue bulletins, reports, and alerts to their agencies. They then work with their police agencies to develop effective strategies and tactics to address crime and disorder. Other duties of crime analysts may include preparing statistics, data queries, or maps on demand; analyzing beat and shift configurations; preparing information for community or court presentations; answering questions from the public and the press; and providing data and information support for a police department's CompStat process.

To see if a crime fits a certain known pattern or a new pattern is often tedious work of crime analysts, detectives or in small departments, police officers or deputies themselves. They must manually shift though piles of paperwork and evidence to predict, anticipate and hopefully prevent crime. The U.S. Department of Justice and the National Institute of Justice recently launched initiatives to support "predictive policing", which is an empirical, data-driven approach. However this work to detect specific patterns of crime committed by an individual or group, (crime series), remains a manual task. Over the past year, MIT doctoral student Tong Wang, Cambridge (Mass.) Police Department CPD Lieutenant Daniel Wagner, CPD crime analyst Rich Sevieri and Assoc. Prof. of Statistics at MIT Sloan School of Management and the co-author of"Learning to Detect Patterns of Crime" Cynthia Rudin have designed a machine learning method called "Series Finder" that can assist police in discovering crime series in a fraction of the time. Series Finder grows a pattern of crime, starting from a seed of two or more crimes. The Cambridge Police Department has one of the oldest crime analysis units in the world and their historical data was used to train Series Finder to detect housebreak patterns. The algorithm tries to construct a modus operandi (MO). The M.O. is a set of habits of a criminal and is a type of behavior used to characterize a pattern. The data of the burglaries include means of entry (front door, window, etc.), day of the week, characteristics of the property (apartment, house), and geographic proximity to other break-ins. Using nine known crime series of burglaries Series Finder recovered most of the crimes within these patterns and also identified nine additional crimes. Machine learning is a tremendous tool for predictive policing. If patterns are identified the police can immediately try to stop them. Without such tools it can take weeks and even years of shifting though databases to discover a pattern. Series Finder provides an important data-driven approach to a very difficult problem in predictive policing. It's the first mathematically principled approach to the automated learning of crime series.Sociodemographics, along with spatial and temporal information, are all aspects that crime analysts look at to understand what's going on in their jurisdiction. Crime analysis employs data mining, crime mapping, statistics, research methods, desktop publishing, charting, presentation skills, critical thinking, and a solid understanding of criminal behavior. In this sense, a crime analyst serves as a combination of an information systems specialist, a statistician, a researcher, a criminologist, a journalist, and a planner for a local police department.

II. WEB CRIME MINING

Web mining - is the application of data mining techniques to discover patterns from the World Wide Web. Web mining can be divided into three different types – Web usage mining, Web content mining and Web structure mining.

Computer crime, or cybercrime, is crime that involves a computer and a network.] The computer may have been used in the commission of a crime, or it may be the target."Offences that are committed against individuals or groups of individuals with a criminal motive to intentionally harm the reputation of the victim or cause physical or mental harm, or loss, to the victim directly or indirectly, using modern telecommunication networks such as Internet (Chat rooms, emails, notice boards and groups) and mobile phones (SMS/MMS)". Such crimes may threaten a nation's security and financial health. Issues surrounding these types of crimes have become high-profile, particularly those surrounding hacking, copyright infringement, child pornography, and child grooming. There are also problems of privacy when confidential information is intercepted or disclosed, lawfully or otherwise. Define cybercrime from the perspective of gender and defined 'cybercrime against women' as "“Crimes targeted against women with a motive to intentionally harm the victim psychologically and physically, using modern telecommunication networks such as internet and mobile phones”.A report (sponsored by McAfee) estimates that the annual damage to the global economy is at $445 billion;[5] however, a Microsoft report shows that such survey-based estimates are "hopelessly flawed" and exaggerate the true losses by orders of magnitude. Approximately $1.5 billion was lost in 2012 to online credit and debit card fraud in the US.

1. The diversity of the information on the multiple pages show similar information in different words or formats based on the diverse authorship of Web pages that make the integration of information from multiple pages as a challenging problem.
2. An association is exist on the significant amount of information of the Web. Hyperlinks are in Web pages across different sites and within a site. Hyperlinks are implicit conveyance of authority to the target pages in across different sites. And hyperlinks serve as information organization mechanisms within a site.
3. The information on the Web is noisy that is comes from two main sources. The first one is that a typical Web page involves many pieces of information for instance the navigation links, main content of the page, copyright notices, advertisements, and privacy policies. Only part of the information is useful for a particular application but the rest is considered noise. For performing a fine-grain, the data mining and Web information analysis, the noise should be removed. The second one is due to the fact that the Web does not have quality control of information, for example, a large amount of information on the Web is of low quality because any one can write everything.
4. The Web is about services for example most commercial Web sites allow the users to perform useful operations at their sites such as paying bills, purchasing products, and filling the forms.
5. The Web pages are dynamic that is the information is changes constantly. Copping the changes and monitoring them is an important issue for many applications.
6. The Web is a virtual society that is not only information, data and services; it also is the organizations, the interactions of people, and automated systems. Any user can communicate with people anywhere in the world easily and express his/her views on anything in Internet blogs, forums and review sites.

III.CRIME DATA MINING TECHNIQUES

Data Clustering : In this section, considerations and challenges of leveraging classic hierarchical and partitioning clustering techniques in intelligent crime analysis has been discussed. Subsequently, a proposed approach for crime data clustering is represented which utilizes SOM neural network in order to overcome other clustering techniques drawbacks. The binary nature of crime behavioral variables may be regarded as a challenge, because committing clustering and analysis process on these types of variables requires some elegance. The binary encoding causes the popular Euclidian distance measure- which is commonly used for continuous types of variables- to be useless. The reason is that behaving binary quantities like continuous quantities can lead to misleading results in clustering process [8]. Some other distance functions should be leveraged which are specific for achieving the similarity between binary data objects. These functions calculate the dissimilarity between two objects as their corresponding distance. The distance (dissimilarity) between two objects can be calculated by equation .

$$D\ (i\ ,\ j)= 1- S\ (i\ ,\ j)\ (1)$$

Where S and D correspondingly represent functions of similarity and dissimilarity between two binary sequences i and j. In order to calculate the similarity, assuming that both sequences are equal in length, we need to define M.R. Keyvanpour et al. / Procedia Computer Science 3 (2011) 872–880 875 MohammadReza Keyvanpour / Procedia Computer Science 00 (2010) 000–000 variables a, b, c and d as follows: a represents the number of corresponding bits with value 1 in both bit sequences. b represents the number of bits which equal 1 in the first bit sequence but equal 0 in the second bit sequence. c represents the number of bits which equal 0 in the first bit sequence but equal 1 in the second bit sequence. Finally, d represents the number of corresponding bits with value 0 in both bit sequences. Table 2. Different modes of bit comparison Respective variable Bit value in the first sequence Bit value in the second sequence a1 1 b1 0 c0 1 d0 0 According to the above definitions the similarity between bit sequences i and j can be measured by the following equations [9]: - Simple Match Coefficient:

$s(i,j) = (p+s) / (p+q+r+s)$ - Rao's Coefficient:
$s(i,j) = p / (p+q+r+s)$ - Jaccard Coefficient: $s(i,j) = p / (p+q+r)$

It is also notable that popular classic ţ-means algorithm cannot be used for clustering binary data objects. The reason is that, classic ţ-means' calculated centroids are not binary. To overcome the problem, using ţ-medoids partitioning clustering method is suggested, in which one of the binary objects in a cluster represents the center of that cluster. Complementarily, if hierarchical clustering method is chosen in order to cluster binary data objects, it will be impossible to use Average-link or centrroid distance methods as inter-cluster distance measuring strategy. Instead, we can use .Single-link or complete-link distance measures. It is also worthy to know that because different types of hierarchical clustering methods have time complexity of O(n2) and O(n2 logn), it is not lucrative to use hierarchical clustering with large amount of high-dimensional crime data.Association rule mining : It determines frequently occurring item sets in a database and offerings some patterns as rules that been used in network intrusion detection to develop the connection rules from users' interaction history. Investigators also can apply this technique to network intruders' profiles to help detect potential future network attacks. Similar to association rule mining, sequential pattern mining finds frequently occurring sequences of items over a set of transactions that occurred at different times. In network intrusion detection, this approach can identify intrusion patterns among time-stamped data. Showing hidden patterns benefits crime analysis, but to obtain meaningful results requires rich and highly structured data.

Deviation detection utilizes the particular measures to study data that differs noticeably from the rest of the data. Also called outlier detection, investigators can apply this technique to fraud detection, network intrusion detection, and other crime analyses. However, such activities can sometimes appear to be normal, making it difficult to identify outliers.

Classification : Finds mutual properties between various crime entities and arranges them into predefined classes that have been applied for identifying the source of email spamming according to the sender's structural features and linguistic patterns . Often used to predict crime trends, classification can reduce the time required to identify crime entities. However, the technique requires a predefined classification scheme. Classification also requires reasonably complete training and testing data because a high degree of missing data would limit prediction accuracy. String comparator techniques that show the relation the textual fields in pairs of database records and calculate the correspondence among the records that can detect deceptive information in criminal records for instance the name and address. the researchers can utilize string comparators to evaluate textual data that often need intensive computation. String comparison is the interesting field for computer scientists that whether string matching or string distance measures. Levenshtein define a usual measure Detecting Suspicion Information on the Web using Crime of similarity between two strings as “edit distance” so, the minimum number of, deletions, single character insertions, and substitutions need to transform one string into the other.

IV CONCLUSION

The majority of digital evidence is collected from textual data such as blogs, as e-mails, web pages, text documents and chat logs. The researcher uses some search tools to explore and extract the useful information from the text because the nature of textual data is unstructured and then for further investigation, enter the appropriate pieces into a well-structured database manually which will be boring and error prone.Therefore, the investigators expertise and experience is very important in search and the quality of an analysis. If a criminal hide some essential information, it may be missed.In this review all preliminary concepts such as Web Mining, Criminal Identities and Crime Data Mining Techniques are described. The vision of the Web Mining is to provide a Web where all published material is understandable by software agents. Moreover, Data Mining defined as the process of discovering useful patterns or knowledge from data sources, e.g., databases, texts, images, the Web, etc.Web mining aims to discover useful information or knowledge from the Web hyperlink structure, page content and usage data. Inspection of files involves searching content for information that can be used as evidence or that can lead to other sources of information that may assist the investigation process and analysis of the retrieved information. It is typically up to the investigator on what and how to search for evidence, depending on the case.

Therefore, we evaluated State-of-the-Art approaches for extracting useful information by means of Data Mining, in order to find crime hot spots out and predict crime trends for them using crime data mining techniques.A web page involving a crime can be thought of as a chain of actions with a series of background attributes. Thus, we can analyze web information from the perspective of events and apply some research results related to events to solve the problem of web crime mining. An event is identified by event triggers, is associated with participants, time, location, et al., and is a larger semantic unit compared with a concept.

There is an intrinsic link between events. It is a new attempt to apply the semantic analysis technology of events to mine web crime information on the web. The majority of digital evidence is collected from textual data such as blogs, as e-mails, web pages, text documents and chat logs. The researcher uses some search tools to explore and extract the useful information from the text because the nature of textual data is unstructured and then for further investigation, enter the appropriate pieces into a well-structured database manually which will be boring and error prone. Therefore, the investigators expertise and experience is very important in search and the quality of an analysis. If a criminal hide some essential information, it may be missed.Mining defined as the process of discovering useful patterns or knowledge from data sources, e.g., databases, texts, images, the Web, etc. Web mining aims to discover useful information or knowledge from the Web hyperlink structure, page content, and usage data.Inspection of files involves searching content for information that can be used as evidence or that can lead to other sources of information that may assist the investigation process and analysis of the retrieved information. It is typically up to the investigator on what and how to search for evidence, depending on the case. Therefore, we evaluated State-of-the-Art approaches for extracting useful information by means of Data Mining, in order to find crime hot spots out and predict crime trends for them using crime data mining techniques.

REFERENCES

1. Fayyad, U.M., and Uthurusamy, R. (Aug. 2002). Evolving Data Mining into Solutions for Insights. Comm. ACM. 28-31.

2. Hosseinkhani, J., Ibrahim, S., Chuprat, S. and Hosseinkhani, N.J. Web Crime Mining by Means of Data Mining Techniques, Research Journal of Applied Sciences, Engineering & Technology (Print ISSN: 2040-7459 Online ISSN: 2040-7467), 2013.

3. Senator, T. et al., (1995). “The FinCEN Artificial Intelligence System: Identifying

Potential Money Laundering from Reports of Large Cash Transactions", AI Magazine, vol.16, no. 4, pp. 21-39.

4. Levenshtein, V.L., 1966. Binary codes capable of correcting deletions, insertions and reversals. Soviet Phys. Doklady, 10: 707-710.

5. Vel, O. de., A. Anderson, M. Corney and G. Mohay, 2001. Mining e-mail content for author identification forensics. SIGMOD Record, 30(4): 55-64.

6. Agrawal, R. and R. Srikant, 1995. Mining sequentiel motifs. Proceeding of the 11th Intrnational Conference on Data Engineering.

7. Chau, M., J.J. Xu and H. Chen, 2007. Extracting meaningful entities from police narrative reports. Proceeding of the National Conference on Digital Government Research. Digital Government Research Center, pp: 271-275.

8. Han, J. and M. Kamber, 2009. Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco.