http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 5 (1): 57-63, 2006
ISSN 1812-5638
© 2006 Asian Network for Scientific Information

Applying Data Mining Techniques in Intrusion Detection System on Web and Analysis of Web Usage

[1]Alaa H. Al-Hamami, [2]Mohammad Ala'a Al-Hamami and [2]Soukaena Hassan Hasheem
[1]Al- Ahliyya Amman University, Amman, Jordan
[2]Department of Computer Science, Al-Rafidain University College, Baghdad, Iraq

Abstract: This research concentrates on a particular aspect, which is to detect the intrusion on the web by applying Data Mining Technique especially by association analysis algorithm on the web log files. This algorithm depends on the facts that define the intrusion. As known the association algorithm first extracted the large itemsets that represent the super facts to build the association analysis for the intrusion. After building the intrusion detection scheme, all the transaction presented in the intrusion was deleted. Analysis the web log data file was again analyzed for the web usage, to study the visitor tracking without tracking the intruder visitors. According to this study the server configurations and all the services were improved.

Key words: Intrusion detection, data mining, association analysis, web usage

INTRODUCTION

An attack is generally unwanted intrusion. Attack strategies often concentrate on vulnerabilities (also called holes or backdoors) of specific operating system or hardware of networks. There are two general types of attacks: Passive Attacks, where the intruder does not interfere with the system or attempt to cause any damage to it. But simply monitors private data, usually in transit, if necessary make cryptanalysis to attempt to break any encryption in use, anyway it is also dangerous. Active Attacks, where intruder does interfere with data or resources in the targeted system or network. Such attacks include masquerades by address spoofing, modification or fabrication of files or messages and the use of the available resource[1,2].

Computer criminals are kind of people that penetrate security of systems. They defined as follow: Hackers are people who gain unauthorized access to computer or telecommunications systems, often just for challenge of it. Hacker enjoys working with computers and spends numerous hours writing program to penetrate secure systems. Cracker is a person who breaks into or otherwise violates the system integrity of remote machine, with malicious intend. Crackers, having gained unauthorized access, destroy vital data, deny legitimate users service, or basically cause problems for their target[3].

With data mining the techniques of the search engines and visitor tracking called web mining. The important task for web mining is web usage mining, which mines Web log records to discover user access patterns of Web pages. Analyzing and exploring regularities in Web log records identify potential customers for electronic commerce, enhance the quality and delivery of Internet information services to the end user and improve Web server system performance. A web server usually registers a (Web) log entry, or Web log entry, for every access of a Web page. It includes the requested URL, the IP address from which the request is originated and a time stamp. For Web-based e-commerce servers, a huge number of Web access log records are being collected. Popular Web sites may register the Web log records in order of hundreds of megabytes every day[4].

Since Web log data provide information about what kind of users will access what kind of Web pages, Web log information can be integrated with Web content and Web linkage structure mining. This information will help Web page ranking, Web document classification and the construction of a multilayered Web information base as well[5].

DATA MINING AND KNOWLEDGE DISCOVERY

With the enormous amount of data stored in files, databases and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially

Corresponding Author: Dr. Alaa H. Al-Hamami, Professor, Al- Ahliyya Amman University, Amman, Jordan

useful information from data in databases. While data mining and Knowledge Discovery in Databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process[6].

In principle, data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. However, algorithms and approaches may differ when applied to different types of data. Indeed, the challenges presented by different types of data vary significantly. Data mining is being put into use and studied for databases, including relational databases, object-relational databases and object oriented databases, data warehouses, transactional databases, unstructured and semi structured repositories such as the World Wide Web, advanced databases such as spatial databases, multimedia databases, time-series databases and textual databases and even flat files.

World Wide Web is the most heterogeneous and dynamic repository available. A very large number of authors and publishers are continuously contributing to its growth and metamorphosis and a massive number of users are accessing its resources daily. Data in the World Wide Web is organized in inter-connected documents. These documents can be text, audio, video, raw data and even applications. Conceptually, the World Wide Web is comprised of three major components: the content of the Web, which encompasses documents available, the structure of the Web, which covers the hyperlinks and the relationships between documents; and the usage of the web, describing how and when the resources are accessed. A fourth dimension can be added relating the dynamic nature or evolution of the documents. Data mining in the World Wide Web, or web mining, tries to address all these issues and is often divided into web content mining, web structure mining and web usage mining[7].

WHAT CAN BE DISCOVERED ?

The kinds of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties of the existing data and predictive data mining tasks that attempt to do predictions based on inference on available data[8]. The data mining functionalities and the variety of knowledge they discover are briefly presented in the following list:

Characterization: Data characterization is a summarization of general features of objects in a target class and produces what is called characteristic rules.

Discrimination: Data discrimination produces what are called discriminant rules and is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class.

Association analysis: Association analysis is the discovery of what are commonly called association rules. It studies the frequency of items occurring together in transactional databases and based on a threshold called support, identifies the frequent item sets. Another threshold, confidence, which is the conditional probability than an item appears in a transaction when another item appears, is used to pinpoint association rules.

Classification: Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects.

Clustering: Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes.

INTRUSION DETECTION TECHNIQUES

There are mainly two types of intrusion detection techniques: Anomaly detection and Misuse detection. Most current intrusion detection systems use one or both of these two approaches: Anomaly detection, for example IDES, tries to determine whether deviation from an established normal behavior profile can be flagged as an intrusion. A profile typically consists of a number of statistical measures on system activities, for example, the CPU usage and the frequency of system commands during a login session (of a user). Deviation from a profile can be computed as the weighted sum of the deviations of the constituent statistical measures. Profiles can be updated periodically (aged) so that shifts of normal behavior are accounted for. Misuse Detection refers to techniques that use patterns of known intrusions (for example, more than three consecutive failed logins within 2 minutes is a penetration attempt) or weak spots of a system (for example, system utilities that have the buffer overflow vulnerabilities) to match and identify intrusions. The sequence of attack actions, the conditions that compromise a system's security, as well as the evidence

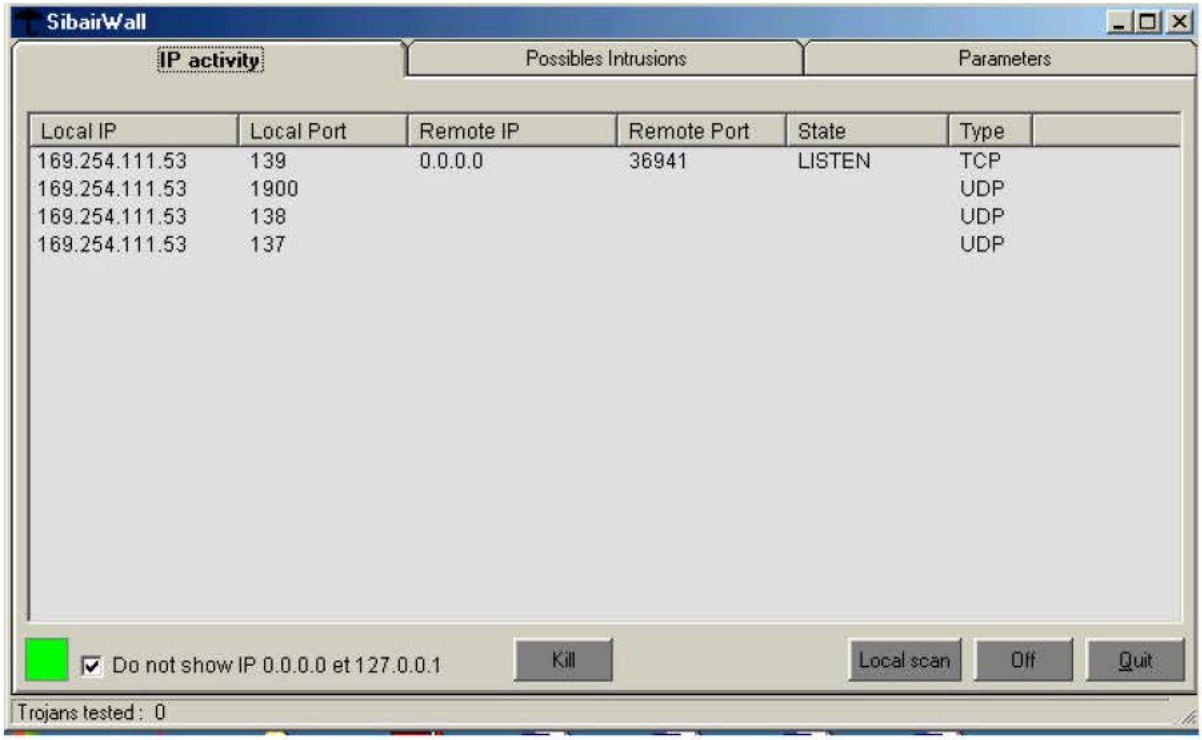

Fig. 1: Web log in/out file

TID	Local IP (A)	Local port (B)	Remote IP (C)	Remote port (D)	State (E)	Type (F)	Time stamp (G)
1	122.22.3.18	139	33.56.233.77	80	Listen	Tcp	2:30
2	122.22.3.18	139	44.56.78.22	50	Listen	Udp	5:50
.							
.							
.							
.							
.							

Fig. 2: The relational Database (D) for the web login/out file

(e.g., damage) left behind by intrusions can be represented by a number of general pattern matching models, for example, NIDES uses rules, STAT uses state transition diagrams and uses Colored Petri nets.

In today's network computing environment, there are multiple penetration points for intrusions to take place. For example, at the network level, intruders can crash a host by simply sending malicious packets to the victim; at the host level, intruders can first login to a system and then illegally obtain a root (superuser) shell. System activities occurring at these different "penetration points" are captured in different audit data sources. However many intrusion detection systems handle only one particular audit data source because of the knowledge- and labor-intensive nature in understanding the system audit data, selecting statistical measures and modeling intrusion patterns.

The large traffic volume in security related (industry groups and underground societies) mailing lists and Web sites suggest that new system security holes and intrusion methods are continuously being discovered. Therefore it is imperative that intrusion detection systems be updated and upgraded frequently. However these maintenance and updates are expensive and difficult because of the current manual, ad hoc approaches.

THE PROPOSED SYSTEM

In this research we describe a data mining framework for constructing intrusion detection models and web usage exactly at the same time. The key ideas are to mine web log record data for consistent and useful patterns of program and user behavior. Also uses the set of relevant system features presented in the patterns to recognize

intrusions in early times, then by eliminate these patterns from the web log data we could apply the web usage with high precision. We propose to use the association rules and frequent itemsets computed from logging data as the basis for guiding an intrusion detection and web usage. The general algorithm and all the details would be explained in the following sections briefly:

The preprocessing stage: In this system we use DM as a basic tool for both of intrusion detection and web usage. This can be done by applying the association analysis algorithm on the web log data. So first preprocess the web log data, this processing include conversion this web log data to a relational database and then apply the association analysis on it. We use the sibairwall program for dealing with the data of web log, (Fig. 1). Then the conversion of the web data log to relational Dtabase (D) was as in Fig. 2.

The basic algorithm: Let A be a set of attributes (local IP (A), local port (B), remote IP (C), remote port (D), state (E), type (F) and time stamp (G)) and I be a set of values on A, called items. Any subset of I is called an itemset. The number of items in an itemset is called its length. Let D be a database with n attributes (columns). Define support (X) as the percentage of transactions (records) in D that contain itemset X. An association rule is the expression $X \rightarrow Y$, c, s. Here, X and Y are itemsets and $X \Phi Y = \cap$. s = support (X U Y) is the support of the rule and c = support(X U Y)/support (X) is the confidence.

The first part (intrusion detection): After conversion of web data file to relational database is completed, this D must be submitted to DM tool, so first we make some transformation on the attribute values in D to be good area for the DM algorithm. This mean convert the D to binary relational database according to the information feeding that describe all possible intrusion; for example if the attributes of the D are have the following values:

A = 120.77.0.10
B = 80
C = 33.40.200.1
D = 67
E = listen
F = tcp
G = 2:22

So this transaction will be CDEF in the binary representation, while the CDEF represent the first face of intrusion. Continue to represent all the transaction according to the scheme of the intrusion; this scheme organized according to the web administration[1].

Now the data become good and clear area to built the first system, intrusion detection system, by using the apriori data mining algorithm as a tool to extract the patterns of the intrusion and built the basic classification rules to classify all the future different patterns.

We implemented the association rules algorithm following the ideas of Apriori, the apriori algorithm declared completely[8].

After the implementation of aprior DM algorithm on the binary D, the results were the patterns of the intrusion and these patterns are declared by the association rules which represented the final output of the above algorithm. Finally these patterns considered as a basic scheme for classification of the future transactions or records entered to the web log files.

The second part (web usage): After the first part is completed, the second part begins with eliminate all the transactions that represented the intrusion from D and then deal with that new D, which is represent clear web log data without intrusion data. The dealing is to analyze the data of web log included in the new D for study and trace the visitors tracking and according to these observations the web administrators would improve their servers and all their services. The analysis is done by the apriori DM algorithm, which is explained in the previous section.

Since we look for correlation among values of different features (attributes) and the (preprocessed) log data usually has multiple columns of features, each with a large number of possible values, we do not convert the data into a binary database.

For example, visitor tracking can give the time periods patterns for which visitors access your site:
For the 24 h covered on 13-April-99:

5–6 am = 23visitors enter by UDP protocol on 80 port from EUR countries only for games.
6– 7 am = 35 visitors from different countries for email.
7–8 am = 42 visitors from Arabian for newsgroup.
8–9 am = 69 visitors almost those are merchants dealing only with e-commerce.
9–10 am = 105 visitors those using searching techniques for some of their interested subjects.
10–11 am = 323 visitors shopping.

From these discovered patterns by DM algorithm the web site administration would be able to depend on these patterns for reconfigure their site servers according to the way preferred by visitors and almost make their predictions about the track of visitors in the coming times.

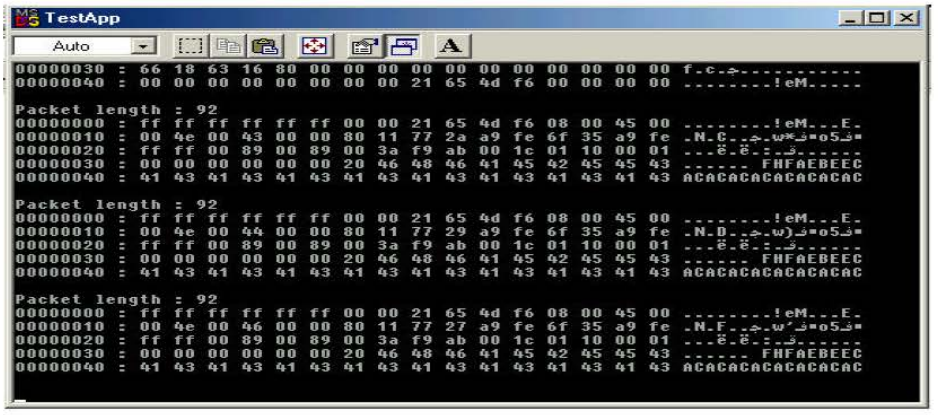

Fig. 3: Packet capture software

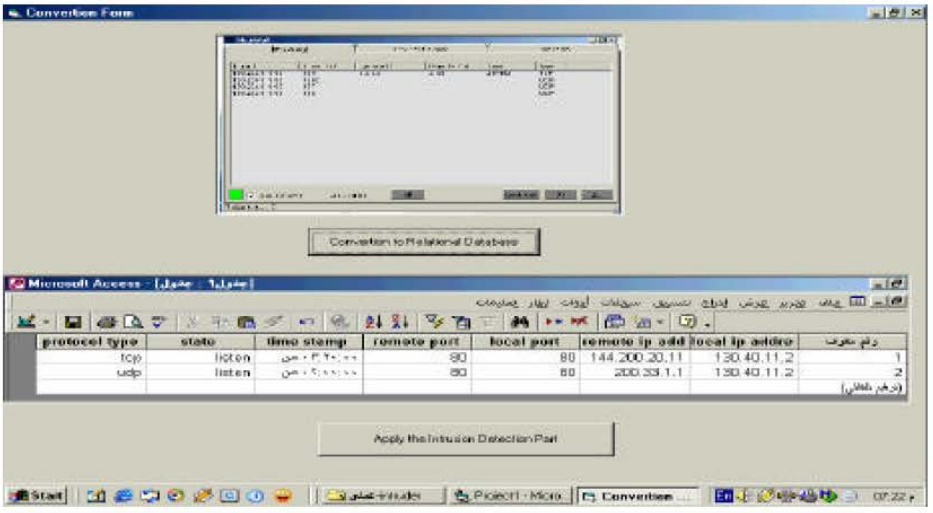

Fig. 4: The conversion form

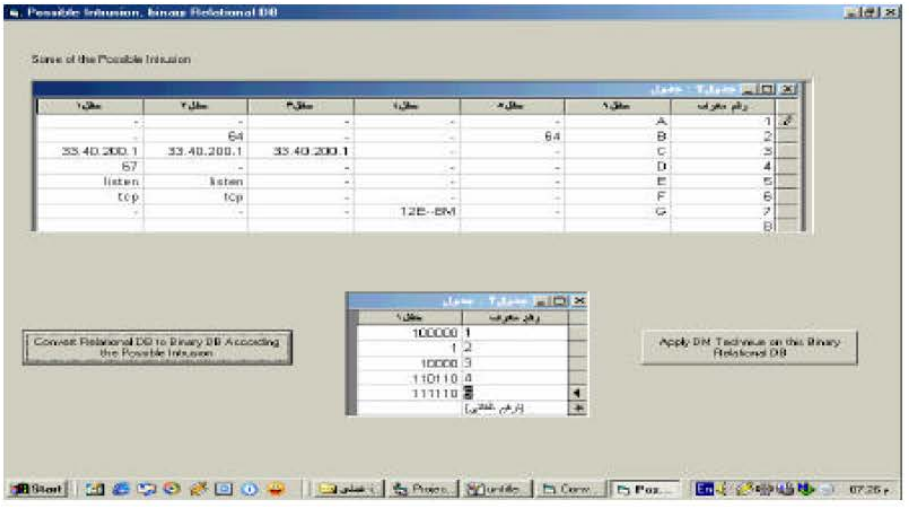

Fig. 5: Some possible intrusion and the binary dbase

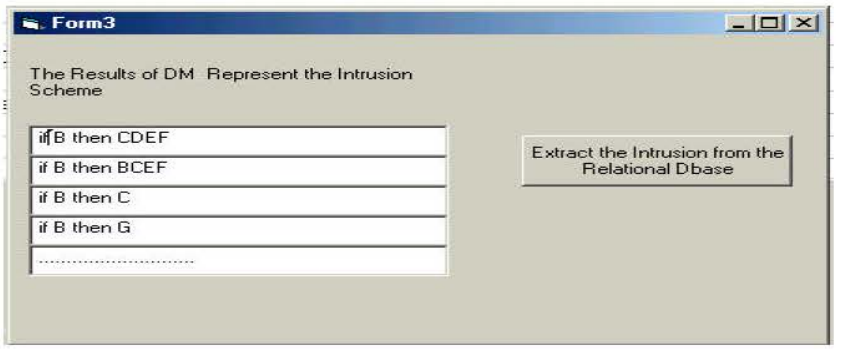

Fig. 6: The results of intrusion according DM technique

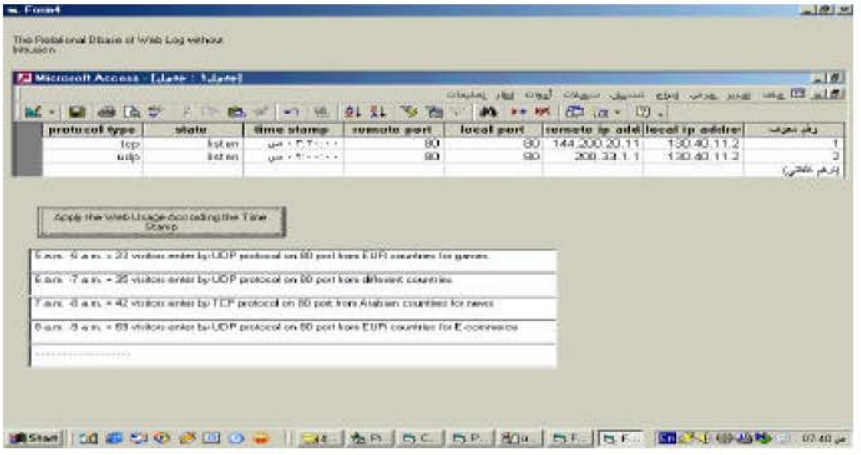

Fig. 7: Relational DB without intrusion with the web usage result

THE IMPLEMENTATION

To display the idea of this research in more clearly phase we would present the real implementation of that system as follow:

First, we must explain that this system would work as an offline system, because it collects the web log information and submits these information to this system to detect the scheme of intrusion on the web and then study web tracking to improve the configuration of that web server according to what it preferred by the visitor.

The application program sibairwall collects the most important information about the visitor by using the packet capture software as shown in Fig. 3.

The software responsible is to take the hexadecimal values of local address, local port, remote address, remote port, state and time stamp and protocol type from the packet. This information will be shown in understandable information for the administration. After the collection of data was completed these data would be stored in to relational database as shown in Fig. 4.

Figure 4 shows two commands the first, when it activated would display the relational dbase which has all the information collected by the sibairwall software. The activation of the second, would display the contents of Fig. 5.

Figure 5 shows two commands, when the first one is activated the relational dbase would be converted to binary dbase according to the possible facts of intrusion. When the second command is activated would display the contents of Fig. 6.

In Fig. 6 there is only one command, by activation it would extract the transactions will represent the intrusion from the relational dbase and would display the new relational dbase as in Fig. 7.

Figure 7 contains one command, when it activated the program would apply the DM association analysis tool on that relational dbase to extract the visitors patterns according to the time stamp field.

CONCLUSIONS

In context with the results of the present study it can be concluded that:

- Using intrusion detection system with web log data would perform the process of the intrusion detection in very early time, rather than in audit data.
- It is very useful to convert the web data to relational binary database and feeding this database with the super facts of intrusion. So applying the association analysis DM algorithm on web log data would extract the patterns that represent the intrusion and built the basic intrusion scheme to classify the patterns which come in future.
- Mining of the web usage is a great help for support the reputation of the site, this by tracking the visitors and reconfigure the server organization according to it.
- Building web usage on the web log data after applying the intrusion detection system and extracting the intrusion transactions from these data will present high Web usage appreciation.

REFERENCES

1. Escamilla, T., 1988. Intrusion Detection Network Security Beyond the Firewall. Published by John Wiley, Sons, Inc.
2. Banks, M.A., 2001. Computer Security. SYBEX, Inc.
3. Unix Propeller Head, 1997. Maximum Security: A Hacker's Guide to Protecting Your Internet Site and Networks. Macmillan Computer Publishing, Sams Net.
4. Wolder, B., 2001. Internet/Intranet Security NSS Group.
5. Comer, D.E., 2001. Computer Network and Internets with Internet Application. 3rd Edn., Prentice-Hall, Inc.
6. Chen, M.S., J. Han and P.S. Yu, 1996. Data mining: An overview from a database perspective. IEEE Trans. Knowledge and Data Eng., 8: 866-883.
7. Fayyad, U.M., G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, 1996. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press
8. Han, J. and M. Kamber, 2000. Data Mining: Concepts and Techniques. Morgan Kaufmann.