www.ijaret.org Vol. 1, Issue X, Nov.2013
ISSN 2320-6802

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS…..

A REVIEW ON DETECTING VENOMOUS FILES IN IDS USING DATA MINING

Sukhleen[1], Sheveta Vashisth[2]

[1]Student,
Lovely Professional University
Phagwara, Pin no. 144402
sukh.leen@yahoo.co.in

[2]Asst. Professor
Lovely Professional University
Phagwara, Pin no. 144402
sheveta.16856@lpu.co.in

Abstract: *Data mining is an ambiguous term that has been used to refer to the process of finding interesting information in large repositories of data. IDS are known as the intrusion detection system. In reality it is not possible to prevent security breaches completely using the existing security technologies. The intrusion detection plays an important role in network security and information system. In our purposed work, we are going to use the IDS in data mining; it helps to easily find out the various vulnerabilities and attacks in data mining. By the use of IDS, we can enhance the performance of the data mining.*

Keywords: *data mining, intrusion detection, attacks, security, performance.*

1. INTRODUCTION

Data mining is the process to extract the useful information from the large data set. Data mining can be applied to any type of data ranging from weather forecasting, electric load prediction, product design, etc. Data mining is used to automate the detections of relevant patterns in the database. Data mining is an ambiguous term that has been used to refer to the process of finding interesting information in large repositories of data. More precisely, the term refers to the application of special algorithms in a process built upon sound principles from numerous disciplines including statistics, artificial intelligence, machine learning, database science, and information retrieval [2]. Data mining algorithms are utilized in the process of pursuits variously called data mining, knowledge mining, data driven discovery, and deductive learning. Data mining techniques can be performed on a wide variety of data types including databases, text, spatial data, temporal data, images, and other complex data [3].

Data mining tasks can be classified into two categories namely descriptive mining & predictive mining. The descriptive mining techniques such as clustering, Association, Sequential Pattern discovery, is used to find human interpretable patterns that describe the data. The predictive mining techniques like classification, Regression, and Deviation detection, etc., are used to predict unknown or future values of other variables.

Data Mining Architecture: In the data mining, data is stored in the databases. Data mining is the process of discover or extracting useful knowledge from large amounts of data stored in multiple data sources such as file systems, databases, data warehouses etc.[1] This knowledge contributes a lot of benefits to business strategies, scientific, medical research, governments and individual. There are basic four types of architectures are presented in the data mining as:

- **No-coupling**

In the no coupling architecture, data mining system does not use any kind of functionality of the database. The no coupling data mining system retrieves data from a particular data sources. These data sources include file system, processes data using major data mining algorithms. The no coupling architecture is considered as a poor architecture for data mining system, because it is used for simple data mining processes.

- **Loose Coupling**

In the loose coupling architecture, data mining system retrieve the data from the database. This architecture is used for memory based data mining system that does not require high scalability and high performance.

- **Semi-tight Coupling**

In semi tight coupling architecture, It does not link to the database rather it uses the several features of the data mining.

With the help of these features it performs some data mining tasks including sorting, indexing, aggregation etc. In this architecture, intermediate result can be stored in database for better performance.

- **Tight Coupling**

In tight coupling architecture, database is treated as an information retrieval component of data mining system using integration. The database features are used to perform data mining tasks. This architecture provides system scalability, high performance and integrated information.

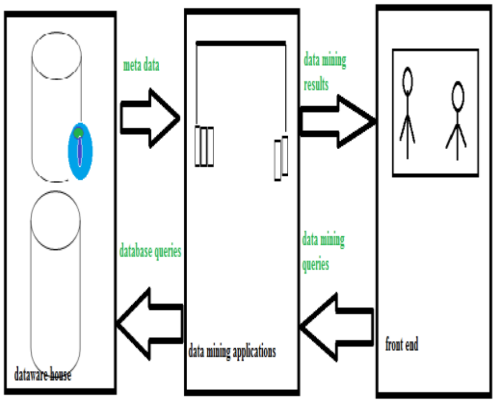

Figure 1: Tight coupling architecture

There are three tiers in the tight coupling data mining architecture:

- **Data layer**: the first layer of the tight coupling architecture is data layer. It can be defined as database or data warehouse systems. This layer is an interface for all data sources.
- **Data mining application layer**: this layer is used to retrieve data from database. Some transformation routine can be performed here to transform data into desired format.
- **Front end layer: this layer pro**vides friendly user interface for end user to interact with data mining system.

Types of Intrusion Detection System

Current IDSs fall into two categories:

- **Network Based IDS**

Because they only scrutinize network traffic [4] NIDS do not benefit from running on the host. As a result, they are often run on dedicated machines that observe the network flows, sometimes in conjunction with a firewall. In this case, they are not affected by security vulnerabilities on the machines they are monitoring. An NIDS monitors network traffic and cannot see the activity going on inside a computer itself. To monitor the activities within a computer system, a company would need to implement a host based IDS. An authorized user of the system may be able to set up an encrypted channel when accessing the machine remotely.

- **Host Based IDS**

HIDS have an ideal vantage point. Because an HIDS runs on the machine it monitors, it can theoretically observe and log any event occurring on the machine. However, the complexity of current operating system often makes it difficult to observe any event. There are certain difficulties faced by security tools that rely on system calls interposition to monitor a host. In addition to shortcomings resulting from an incomplete understanding of the operating system, race conditions in the operating system make the implementation of such tools delicate. HIDSs are also confirmed with difficulties arrived from arising from potential tampering by the attacker.[2]

2. LITERATURE SURVEY

M. Chandrashekhar, K. Raghuveer, (2012) represents a paper," **Performance evaluation of data clustering techniques using KDD Cup-99 Intrusion detection data set"** in this paper author discussed about the intrusion detection using the clustering techniques. The intrusion detection system is used to detect the various attacks against the computer network. For the intrusion detection many techniques are used.[5] Intrusion detection is the process of observing and analysing the events taking place in a computer system. Intrusion detection gives data mining a opportunity to make several important contributions to the field of intrusion detection. Intrusion detection is an essential component of layered computer security mechanism. It requires accurate and efficient models for analyzing a large amount of system and network audit data. Data mining techniques make it capable to search large quantity of data for distinctive rules and patterns. Here author discuss the clustering technique for IDS, in this they use this technique to find out the various attacks or vulnerabilities presents in case of data mining. A clustering technique partitions a data set into several groups, within a group is larger than amongst groups. In this paper four clustering techniques are used. These techniques are: k-means cluster ring, fuzzy c-means clustering, Mountain clustering, and Subtractive-clustering. These techniques are implemented and tested against KDD cup-99 data set.

Manne suneetha et.al, (2011) represents a paper **"Clustering of Web Search Results is using Suffix Tree Algorithm and Avoidance of Repetition of same Images in Search Results using L-Point Comparison Algorithm".** In this paper author describe the clustering in dta mining. Clustering is a search technique, it deals with the grouping the number of similar kinds of data, which is used to retrieved the results that are organized into meaningful groups. In this paper, the suffix tree based clustering is used. [6] In this paper author organized the web search results into clusters facilitating quick browsing options to the browser providing an excellent interface to results precisely. Suffix tree clustering produces more accurate and informative grouped results. In this paper author's main concern is about the image searching. In this the problem of image repetition is occurring, while searching the image. This problem can be solved by using the L point comparison algorithm. With the help of this project user can easily access the results in the form of clusters. With the help of this, user can easily browse the relevant cluster and the repetition of the images is also less.

Wang Pu, Wang Jun-qing, (2011) represents a paper **"Intrusion Detection System with the Data Mining Technologies".** In this paper author discussed the various issues of IDS and the data mining. As intrusion detection becomes an efficient tool for detecting network attacks. Data mining is use to mine the useful information from a large set of information. The information may be noisy, fuzzy or redundant. [7] The intrusion detection system allows network administrator to detect policy violations. Intrusion is an action that tries to destroy data confidentiality, data integrality, and data availability of network information. The purpose of Intrusion Detection Systems is to design the computer system in such a way so that it can easily detect attacks against computer systems over insecure networks. Existing IDS systems can be divided into two categories: anomaly detection and misuse detection or signature detection. Anomaly detection is also known as Behavior detection. It is an approach to detect intrusions by first learning the characteristics of normal activity of users. Then the system uses such characteristics to judge whether the user's activity is normal or not. Misuse detection systems are the approach that tries to match user activity to stored signatures of known exploits or attacks. In this paper, author presents the whole techniques of the IDS with data mining approaches in details.

Norouzian M.R et al (2011), represents a paper "**Classifying Attacks in a Network Intrusion Detection System Based on Artificial Neural Networks**" defined Multi- Layer Perceptron (MLP) for implementing & designing the system to detect the attacks & classifying them in six groups with two hidden layers of neurons in the neural networks.[9] Host based intrusion detection is used to trace system calls. This system does not exactly need to know the program codes of each process. Normal & intrusive behavior are collected through system call & analysis is done through data mining & fuzzy technique.

FAN Ya-qin et .al, (2010), represents a paper , **"Data Mining Based Intrusion Detection System in VPN Application".** In this paper VPN, i.e. virtual private network, as the technology changes day by the day. In this technologically developing era, Internet has become a indispensable role in our daily life. [10] The Internet virus also has attacked our work and life, which has brought about inconvenience and loss. Sometime we can loss our confidential data, so to solve this problem author purposed various algorithms like association regulation data excavates. This algorithm helps to detect the system.

3. PURPOSED WORK

In reality it is not possible to prevent security breaches completely using the existing security technologies. The intrusion detection plays an important role in network security and information system. However, many current intrusion detection systems (IDSs) are signature based systems. The signature based IDS also known as misuse detection looks for a specific signature to match, and identify an intrusion. When the signatures or patterns are provided, they can detect all known attack patterns, but there are some problems for unknown attacks. The rate of false positives is very low but these types of systems are poor at detecting new attacks, variation of known attacks or attacks that act as normal behavior. In this dissertation we use the suffix tree for detecting the attacks in IDS. It will help intrusion detection system for detection of new attacks. KDD99 dataset is used as the training data set.

4. METHODOLOGY

Data Mining: data mining is a technique to mine the useful or important information from the large information. It is a process used by companies to turn raw data into useful information. Data mining depends on effective data collection and warehousing as well as computer processing.

Intrusion Detection system: intrusion detection system is act as the firewall between the user and the databases. There are several files saved in our databases. When a user wants any information, it fetches from the database. If the data is safe i.e. Information does not have the harmful files, it fetches directly by the user. Otherwise it will not fetch the information as:

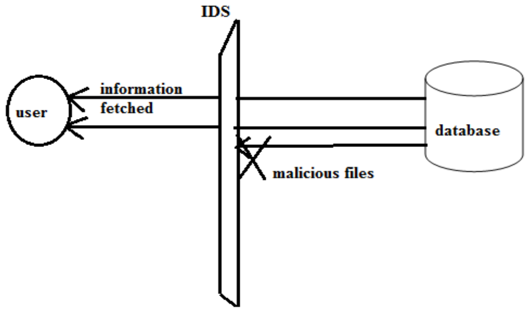

Figure 2: IDS detecting malicious files

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS…..

In the figure 2, IDS is shown. It detects the malicious files from the database.

IDS Using Data Mining: As we know that, data mining is use to discover the information. Here in this case, we use the data mining to discover the malicious information from the data mining, as:

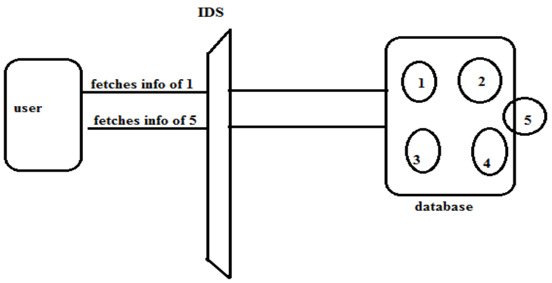

Figure 3: Fetches the malicious files

In the figure 3, the user fetches the information from the database. There are several information are saved in the database. Firstly user fetches the information of DB 1. The information is also stored in the IDS, so it detect it an send it to user. After that a third party add some malicious file, which are having the harmful extensions like, .mat, .bat, the IDS does not capable to detect the malicious files and when user fetches the file it directly sends it to the user. It is the big problem. To solve this problem, in our proposed methodology we are going to use Prefix searching technique with clustering. Till date the searching use in data mining is a suffix searching. In the suffix searching, the algorithm starts mining from first letter of the sentence, which consumes time and decrease performance. Now we are going to use prefix tree, which is totally opposite to suffix and starts searching from last letter. Here clustering is used to make group of similar data. As:

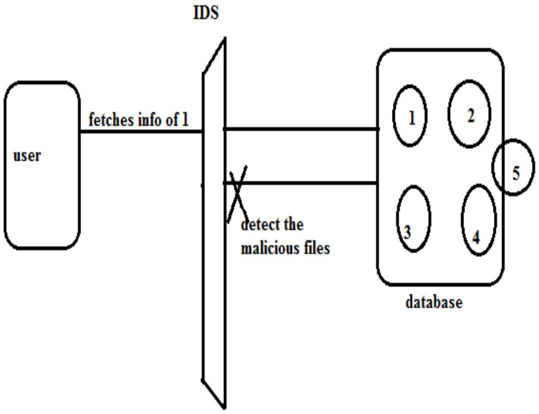

Figure 4: Detection of the harmful files

In the figure 4, the IDS detect the harmful files.

In this case, IDS search the files using the prefix search algo, here it search from the last letter of the extension of the file, if the IDS found the bad extensions it does not read those files and it do not forward that files to user. Hence it protects our system from malicious files in very easy way. It is less time consuming and with the help of this the performance of the system is also enhanced.

REFERENCES

[1]. http://www.zentut.com/data-mining.

[2]. Jiawei Han and. Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kufmann, 2nd edition 2006, 3rd edition 2011.

[3]. Margaret Dunham, (2003) “Data Mining Introductory and Advanced Topics”, ISBN: 0110888923, Prentice Hall.

[4]. Litty Lionel, (2005) “Hypervisor-based Intrusion Detection”, Master of Science Graduate department of computer Science University of Torronto.

[5] A. M. Chandrashekhar, K. Raghuveer, Performance evaluation of data clustering techniques using KDD Cup-99 Intrusion detection data set, International Journal of Information & Network Security (IJINS) Vol.1, No.4, October 2012, pp. 294~305 ISSN: 2089-3299

[6] Manne suneetha, Dr. S Sameen Fatima, Shaik Mohd. Zaheer Pervez, Clustering of Web Search Results using Suffix Tree Algorithm and Avoidance of Repetition of same Images in Search Results using L-Point Comparison Algorithm, PROCEEDINGS OF ICETECT 2011

[7] Wang Pu, Wang Jun qing, Intrusion Detection System with the Data Mining Technologies, 978-1-61284-486-2/111$26.00 ©2011 IEEE

[8] Mohammad Reza Norouzian, Sobhan Merati, Classifying Attacks in a Network Intrusion Detection System Based on Artificial Neural Networks, ISBN 978-89-5519-155-4 868 Feb. 2011 ICACT2011

[9] FAN Ya-qin et.al, Data Mining Based Intrusion Detection System in VPN Application, 2010 WASE International Conference on Information Engineering