IMPROVING SECURITY BY PREDICTING ANOMALY USER THROUGH WEB MINING: A REVIEW

Mahesh Malviya[1], Abhinav Jain[2], Nitesh Gupta[3]
[1&3]Technocrats Institute of Technology, Bhopal, India.
[2]OIST, Bhopal, India
malviya_mahesh@rediffmail.com, gupta.neetesh81@gmail.com

Abstract

The web log data embed much of the user's browsing behavior. Every visit of internet user is recorded in web server log. There are many systems that attempt to predict user navigation on the internet through the use of past behavior, preferences and environmental factors. Ensuring the integrity of computer networks, both in relation to security and with regard to the institutional life of the nation in general, is a growing concern. Security and defense networks, proprietary research, intellectual property, and data based market mechanisms that depend on unimpeded and undistorted access, can all be severely compromised by malicious intrusions. We need to find the best way to protect these systems. In addition we need techniques to detect security breaches. There has been much interest on using data mining for counter-terrorism and cyber security applications. For example, data mining can be used to detect unusual patterns, terrorist activities and fraudulent behavior. In addition data mining can also be used for intrusion detection and malicious code detection. Our current research is focusing extensively for intrusion detection.

Keyword: *web mining, security, intrusion detection*

1 Introduction

In today's business environment almost all companies have their computers connected to the public Internet. As the number of companies with computers and services accessible to the Internet increases, a corresponding increase in the number of attacks against these businesses is also observed. Network based attacks on business computers have been increasing in frequency and severity over the past several years. Consequently, many research efforts have concentrated on network intrusion detection techniques whose goal is to identify such attacks. For example, reports generated from the Computer Emergency Response Team Coordination Center [7] databases illustrate dramatic growth in reported incidents of security breach over the past years. Due to the fact that the numbers of attacks on the global Internet are increasing, it is critical for companies to secure their network and computers. This is especially true for corporations with businesses that are dependent on the Internet. In severe cases of security breach companies may lose business, and eventually become bankrupt, as a result of one successful attack [6].
Security attacks come from different sources. There are external intruders, who are unauthorized users of the machines they attack, and internal intruders, who have permission to access the system with a number of restrictions [1]. Several techniques have been used to prevent unauthorized access to business data; some suitable to prevent the access by external and internal intruders, while others only prevent the access by external intruders. Users' authentication and data encryption are examples of techniques appropriate for both, external and internal intruders, while firewalls can prevent the access by external intruders.

2. RELATED WORK

In recent years, with the widespread use of Intranet and Internet, users have become more and more dependent on the services provided by networked systems where computer programs and potentially sensitive information are kept in (geographically) dispersed systems and exchanged over telecommunication facilities. Distributed systems have emerged to provide the means through which networked systems cooperate to process users' tasks in a seamless and efficient fashion. Such systems

provide tremendous benefits to their users but also raise new challenges specifically, access control and instruction detection [9].

Access Control:

Security access control mechanisms play a key role in the overall structure of any security system. They are responsible for controlling the access permissions to system resources; i.e. determining who has access to which resource and with what type of access. Access control mechanisms rely on the authentication mechanisms to identify
the users and ensuring that they are actually who they claim to be. The most common authentication method used to date is the user ID and password (or PIN number) combination, though other methods, such as bio-metric identification, have been used with varying degrees of success [9].

Numerous studies [12], however, have shown that a large number (if not most) of security breaches are done by unauthorized users impersonating as authorized users (by guessing passwords or stealing them through various means). Other security breaches occur by circumventing the authentication system altogether, by exploiting security holes" in the system. Once the authentication system is broken, the system and the information kept in it become wide open to unauthorized access and malicious usage. The security risk analysis can be applied to any component of the distributed system (e.g. a user, an end-system, a communication link, a LAN, etc.) and would allow the local host to determine the level of security/hostility of the component [10,11].

Intrusion Detection:

As computer attacks become more and more sophisticated, the need to provide effective intrusion detection methods increases. Current best practices for protecting networks from malicious attacks are to deploy a security infrastructure that includes network intrusion detection systems. While those systems are useful for identifying malicious activity in a network, they generally suffer from several major drawbacks: inability to detect distributed
or coordinated attacks, high false alarm rates, and producing large amount of data that is difficult to analyze. A major concern is the high rate of false alarms produced by current Intrusion Detection Systems which undermine
the applicability of such systems. Effective protection of networks from malicious attacks remains a problem in both the research and network administering communities. Monitoring intrusion detection of multiple network systems requires the existence of multiple intrusion detection systems and a framework for integration.

3. LOG FILE

A log file is a file that is used to track the operation performed by any user simply by storing messages generated by an application, service, or an operating system. For example Web servers maintain a log files record for every request made to the server. Log file is generally in American Standard Code for Information Interchange code file format having a .log extension. Log file is also generated by different functioning logs and alert services. Log file has several benefits, which include troubleshooting, security and pro-active system administration. It is a principal part of Security and can be used in the recognition of attacks and intrusions. It is also very helpful in forensic analysis of systems. On the other hand log analysis focuses on learning a user's query behaviour while user navigates a search site. Understanding the user's navigational preferences helps to improve query behaviour. In fact, the knowledge of the most likely user access patterns allows service provides to customize and adapt their sites interface for individual users as well as to improve the site's static structure within the wider hypertext system.

3.1 Log Type

Different kinds of log files are available in our computer system; depending on their characteristics several of them are used for different purposes like security, data retrieval, analysing, Authentication & etc. Several of them are following.

Web log file : A file that records every request and important information about the requests to web server made by users. For example, every time a browser requests a page, an entry is automatically made in this log by the web server, containing information such as the address of the computer on which the browser was running, the time at which the access was made, and the transfer time of the page, etc. Such information is very useful. For example, A Web-administrator is able to use it to judge whether certain pages should be kept in a fast storage medium in order to optimise them. Web log files are also progressively being used to monitor employees' use of the Web: for example, to detect those who visit computer game sites during office hours. The development of software tools for analysing log files has become a major business, and such tools are often referred to as web logging tools.

System log file: Distributed system log file records information about user requests for any resources. Such a log can be configured to contain a huge amount of information about every access to every resource. Security tools can study such log and determine abnormal behaviour, such as a user logging in at unusual times of the day, suggesting that an intruder is masquerading as a regular user.

Firewall log file*:* After configuring the rule-set, installing the firewall and leasing it pass or deny traffic is not good enough. One must continuously monitor firewalls log files. All firewalls log information either locally or to a centralized logging server. So one should review log files daily, if possible first thing in the morning, to see if any suspicious activity occurred overnight, determine whether new IP addresses are trying to probe network, and then write new and stronger firewall rules to block them, and then decide whether to trace the probes and take some sort of management action.

3.2 Log Formats

There are a number of formats, which define the contents of Web, log files: what data is in the file and the order it appears.

Common log format**:** The CLF is a standardized text file format used by web servers when generating log files. Because the format is standardized, the files may be analysed by a variety of analysis programs. It was defined by the national centre for supercomputing applications and is the most popular log file format. Each line in a file stored in the Common Log Format has the following syntax:

"Host ident authuser date request status bytes"

Example:

```
127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326
```

Extended Log Format: ELF is a standardized text file format, like Common Log Format (CLF), which is used by web servers when generating log files, but ELF files provide more information and flexibility.

Example:

```
#Version: 1.0
#Date: 12-Jan-1996 00:00:00
#Fields: time cs-method cs-uri
00:34:23 GET /foo/bar.html
12:21:16 GET /foo/bar.html
12:45:52 GET /foo/bar.html
12:57:34 GET /foo/bar.html
```

3.3. Log Files Significance

Many researchers have proposed and implemented different models which define different measures of system behavior, with an ad hoc presumption that normalcy and anomaly (or illegitimacy) will be accurately manifested in the chosen set of system features that are modeled and measured. Intrusion

detection techniques can be categorized into ***misuse detection***, which uses patterns of well-known attacks or weak spots of the system to identify intrusions; and ***anomaly detection***, which tries to determine whether deviation from the established normal usage patterns can be flagged as intrusions [4]. Misuse detection systems encode and match the sequence of "signature actions" (e.g., change the ownership of a file) of known intrusion scenarios. The main shortcomings of such systems are: known intrusion patterns have to be hand-coded into the system; they are unable to detect any future (unknown) intrusions that have no matched patterns stored in the system. Anomaly detection systems establish normal usage patterns (profiles) using statistical measures on system features, for example, the CPU and I/O activities by a particular user or program.

4. Analysis of LOG files

Individual log files that records activities related to a particular application although useful in many developments contain a lot of data that might not be particularly useful in intrusion prevention systems. However, comparative analysis of different types of log files, coming from different applications run on the same host can reveal useful interrelations that can be used in intrusion prevention systems. The process of defines it as a process dealing with the association, correlation, combining data coming from different sources has been known in the literature as data fusion. We will analyze log files coming from web server installed on one machine, and log file coming from firewall installed on another machine (regular desktop) to find intruder.

5. Proposed Framework

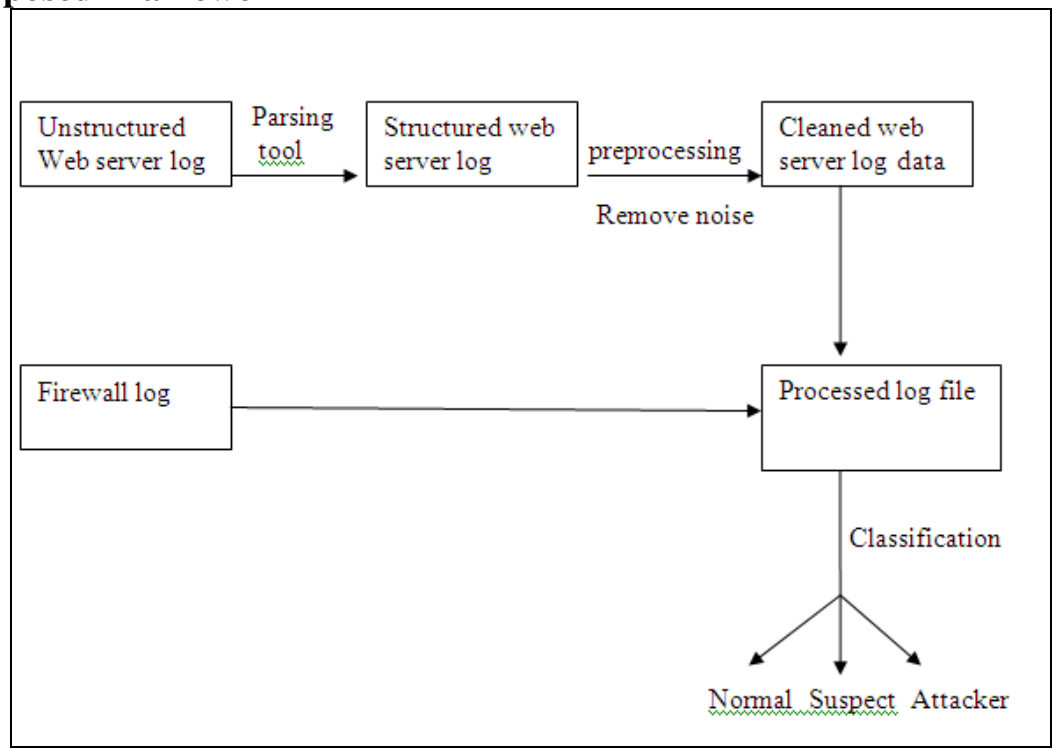

Figure 1. Proposed framework

In the proposed framework unstructured weblog first parsed using parsing tools and produced structured weblog where we can apply data cleaning to remove noise then we find the common entry both in web server log and firewall log. The user whose entry in both firewall and web server and more number of accesses of port 80 become suspicious or attacker. Simple user only visit normal pages and never check whether the connected system is a server or not.

6. Conclusions

Web server logs are mostly captured the behavior of machine not the behavior of end user. Log files provide troubleshooting, security and pro-active system administration that provide significant help in caching suspicious end user. Off line analysis of intersections of log files has allowed us to identify some host IP addresses that most probably belongs to intruders. As those intruders were able to reach our desktop and server that is behind our university firewall, we can provide system administrator with those IP addresses and s/he can set the firewall in such a fashion that those IP will be banned from accessing our network. Intersection of firewall log files coming from different machines can be a source for IP addresses that belong to intruders. Having such information we can create a system that will identify those IP addresses in real-time, and than it will distribute that list to other machines on the LAN to be excluded from their firewall access lists, or it will deliver the list to network firewall.

References:

[1] http://www.webopedia.com/TERM/L/log_file.html
[2] Data Mining – Concept and Techniques by Jiawei Han and Micheline Kamber.
[3] http://www.encyclopedia.com/doc/1O12-systemlog.html
[4] http://en.wikipedia.org/wiki/Common_Log_Format
[5] Tamas Abraham "Event Sequence Mining to Develop Profiles for Computer Forensic Investigation Purposes" Information Networks Division Defence Science and Technology Organization, Australia.
[6] *Buyer's guide for intrusion prevention systems (IPS).*Retrieved June 3, 2004 from http://www.juniper.net/solutions/literature/buryer_guide/710005.pdf
[7] Kazimierz Kowalski, Mohsen Beheshti , Analysis of Log Files Intersections for Security Enhancement.
[8] Bhavani Thuraisingham, Latifur Khan, Mohammad M. Masud, Kevin W. Hamlen Data Mining for Security Applications 2008 IEEE/IFIP International Conference on Embedded and Ubiquitous Computing.
[9] M. Krause and F. H. Tipton, *Handbook of Information Security Management*, CRC Press LLC, Boca Raton, Florida, 1998**.**
[10] A. Berrached, M. Beheshti, A. d. Korvin, and R. Alo, \Intelligent access control in distributed systems using fuzzy relation equations," in *Proceedings of the International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet (SSGRR'00)*, Italy, July 2000.
[11] A. deKorvin, A. Berrached, and M. Beheshti, \Active Access Control in Distributed Systems," in *The 2001 International Conference on Internet Computing (IC'01)*, 2001.

Authors:

Mahesh Malviya: I am Mahesh Malviya student of M. Tech fourth semester in Information Technology branch from Technocrats Institute of Technology Bhopal (MP). I have completed my bachelor degree in 2005 at Mandsaur Institute of Technology ,Mandsaur (MP) with Computer Science & Engineering Branch.and my schooling in my sweet village Khedawad,Post – manglaj ,District –Shajapur (MP) .I have one younger brother,and my interesting area is Data Mining.

Abhinav Jain: I am Abhinav Jain, student of M. Tech Third semester in Computer Science & Engineering branch from Oriental Institute of Science & Technology Bhopal (MP). I have completed my bachelor degree in 2005 at Mandsaur Institute of Technology ,Mandsaur (MP) with Computer Science & Engineering Branch.and my schooling in Ahemdabad (Gujrat) .I have one younger brother,and my interesting area is Data Mining. Web Ming.

Nitesh Gupta: Neetesh Gupta is Ph.D. Research Scholor and he completed B. E. (Computer Science & Engg.Branch) degree from Oriental Institute of Science and Technology / Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal (M.P.)-India in year 2003 and M. Tech. (Computer Science & Engg. Branch) degree from Bansal Institute of Science and Technology/ Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal (M.P.)-India in year 2010. Now he is working as an Asst. Professor in the Department of Information Technology at Technocrat Institute of Technology-Bhopal (M.P.) India since Jan 2005.