International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume **– No.*, ____ 2012 – www.ijais.org

Preprocessing Web Logs for Web Intrusion Detection

Priyanka V. Patil.
M.E. Scholar
Department of computer Engineering
R.C.Patil Institute of Technology, Shirpur, India

Dharmaraj Patil.
Department of Computer Engineering
R.C.Patil Institute of Technology, Shirpur, India

ABSTRACT

Rapid growth in web development provides a way to deliver complex business solution. The increase in web dependency increases web hacking activities. An Intrusion detection has been accepted as a decent security system. It is a process used to identify abnormal activities in a computer system. When user visit website his or her clicks recorded in log file. Log file recorded information about each user. Log file contains unnecessary and noisy data which may affect the results of intrusion detection process.We analyzed log file using preprocessing, preprocessing reduce log file size and increase quality of available data. This paper gives a detailed discussion about the preprocessing web logs. Also Misuse based and Anomaly based Intrusion detection techniques. Intrusion detection plays an important role in addressing security problems of web servers and detecting attacks by monitoring computer system. We propose Web intrusion detection system with Misuse and Anomaly detection mode by using web server access logs.

Keywords

Intrusion Detection, Preprocessing Log File, Host Based, Web Logs.

1. INTRODUCTION

Intrusion Detection is a topic that has recently gathered much interest in the computer security. Intrusion detection is often used as another wall of protection. When user accesses websites are recorded in web logs. Log file contain information about user name, IP address, date, time, bytes transferred, access request. Log files usually contain noisy and unnecessary data. To improve quality of data, log file should be preprocessed. IDS systems are divided into two type Network-based IDS, Host-based. Intrusion detection techniques divided into two categories Misuse based and Anomaly based. Misuse -based detection technique is to find known attack. Anomaly detection technique is good to find known and unknown attacks.

Data Mining, is the process of automatically searching, analyzing, and extracting useful information, from a large volumes of data which is noisy, fuzzy, and random using association rules[1]. Data mining also called as Knowledge Discovery. This paper describes Preprocessing web logs for WebIDS, The Intrusion Detection System isprimary approaches to solve problem of computer security. Our IDS is develop by using Host based IDS. We detect intrusion by using both techniques misuse and anomaly. Our proposed detect web-based attacks like cross site scripting attack and SQL injection attack.

The remainder of this paper is structured as follows, section II discusses literature survey, section III we have given Intrusion Detection, section IV present Methodology. In section V shows experimental results.Section and Conclusion is given section VI.

2. LITERATURE SURVEY

ShaimaaEzzatSalamaet. al. Focused on many works have been devoted to preprocess data in log file for web usage mining. The authors discuss on web log file, Type of web logs, format of log file like NCSA, W3C, IIS. When two different log files converted into one unified XML file, after that preprocessing steps come [2].L. K. Joshila Grace et. al. give a detailed discussion about these log files, their content, their creation or location. Log file located in Web Server, Web Proxy Server and client browser. It also provides the idea of creating an extended log file and learning the user behavior [3].

TheintTheint Aye proposed preprocessing step and to reduce data volume for pattern discovery phase. This paper mainly focuses on data preprocessing step like field extraction and data cleaning algorithms. The field extraction algorithm performs the process of separating fields from the single line of the log file. Data cleaning algorithm eliminates inconsistent or unnecessary items in the analyzed data [4]. Duanyang Zhao et.al. proposed a hybrid IDS,which combines network and host IDS, with anomaly and misuse detection mode, utilizes auditing programs to extract an extensive set of features that describe each network connection or host session. They applies data mining programs like association rule to learn rules that accurately capture the behavior of intrusions and normal activities[5]. Changxin Song, Ke Ma gives idea of applying data mining technology to intrusion detection systems and design data preprocessing module, association analysis module and cluster module respectively Association rule mining and clustering mining are two applications of data mining in intrusion detection systems[6]. C. I. Ezeife et.al.proposed a web intrusion detection ,SesorWebIDS, which applies data mining, anomaly and misuse intrusion detection on web environment for both network and host based intrusion detection[7].

3. INTRUSION DETECTION

Security is becoming a critical part of computer system. Intrusion detection is procedures and system created and operated to detect system intrusion. There are two type of intrusion detection system Network Based IDS, Host Based IDS. Network-based intrusion detection systems monitor network traffic and use raw network packet's content of analyze network, application protocols to identify suspicious activity. Host-based intrusion detection system monitorcharacteristics of a single host and the events occurring within that host, for suspicious activity. Detection of attacks has been performed by applying simple pattern matching techniques to the contents of HTTP requests or by identifying trends in a large set of web-related events. Advantages of HIDS is Monitor who accessed what, Map problem activity to specific user id. Intrusion detection is performed by analyzing one or more input event streams, looking for an attack [8].

Historically, Detection has been achieved by following two different Technique: misuse detection or anomaly detection. Misuse detection systems are equipped with a number of attack descriptions. These descriptions or "signatures" are matched against a stream of audit data.[8] Misuse detection also known as Signatures detection. An anomaly detectoion compare actual usage patterns against established profiles to identify abnormal patterns of activity[8]. It looks for unexpected events. Advantages of Misuse detection systems are good enough to find all known attacks and efficient in performance viz. Drawback of Misuse detection systems is unable to find new intrusions or attacks. Advantages of Anomaly detection system are good enough to find known attacks as well as unknown attacks. Performance viz. anomaly systems are not good. Preprocessing web logs for web intrusion detection are detect SQL Injection and Cross Site Scripting web based attack. SQL injection is a basically a trick to inject SQL command or query as input mainly in the form of the POST or GET method in the web pages.attacks occur when an attacker changes logic of SQL. XSS vulnerability injected script can be all kind of client side such as Java script, ASP script, HTML tag. Research report indicates that more than 80% of the web application are vulnerable to XSS threats.

4. METHODOLOGY

Tremendous uses of web, web log file growing at a faster rate. Log file ranges 1KB to 100MB.We are using web server access log file as input to our system. Web log file resides in web server. Web server have become popular attack target because of their poor security, immediate accessibility and large installation. Standard web server like Apache and IIS. Data from Access logs provides a broad view of web servers and user. General framework of propose system shown in fig.1.

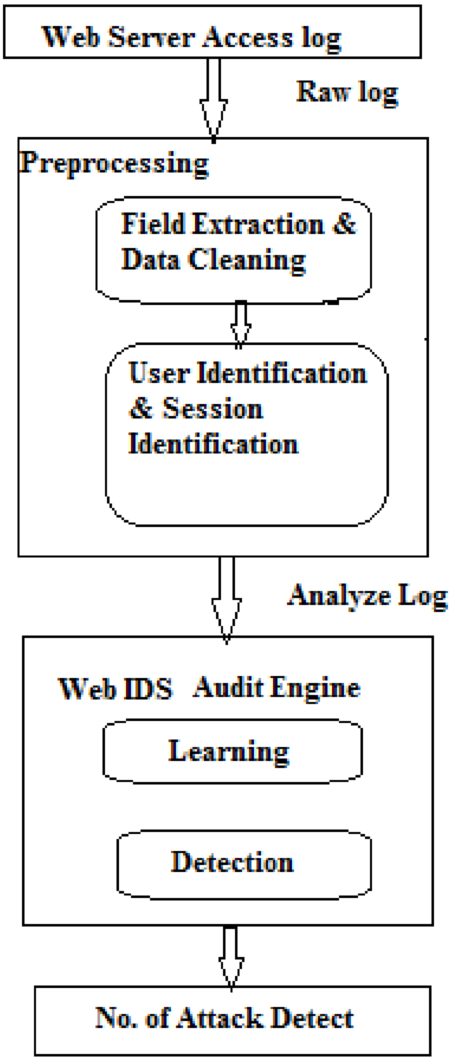
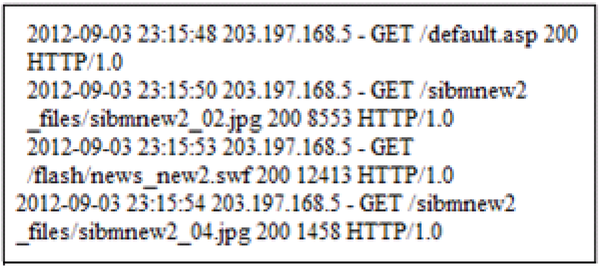


Fig 1.General framework of Propose IDS

Steps of Preprocessing Web Logs for Web Intrusion Detection

Step 1: Data Collection

Our approach has been implemented experimentally on real world data collected from web server log file. Web log file is simple plain text ASCII file. Log file recorded information about each user, so to know user behavior log files are the best source. Log file record shown in fig.2.



```
2012-09-03 23:15:48 203.197.168.5 - GET /default.asp 200
HTTP/1.0
2012-09-03 23:15:50 203.197.168.5 - GET /sibmnew2
_files/sibmnew2_02.jpg 200 8553 HTTP/1.0
2012-09-03 23:15:53 203.197.168.5 - GET
/flash/news_new2.swf 200 12413 HTTP/1.0
2012-09-03 23:15:54 203.197.168.5 - GET /sibmnew2
_files/sibmnew2_04.jpg 200 1458 HTTP/1.0
```

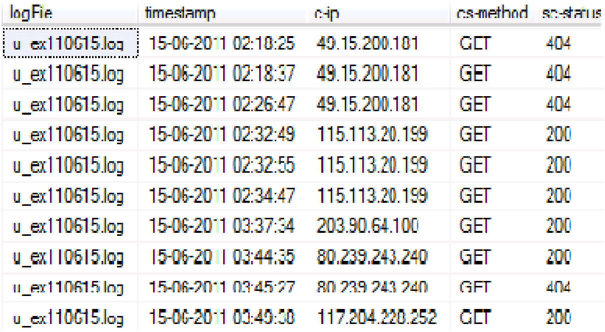
Fig 2. Log file Record

Web server access log records are collected from W3C Extended Log File or NCSA Common Log File. Table 1. Gives detail information about read log record. Statistical report 1 is about no. of records in .css , .jpg, .img etc. shown in Table 2.

Step 2: Field Extraction and Data Cleaning

Field extraction is the first step of preprocessing to read attributes from log file. The log file contains no. of attribute, extract only that field which is needed further. Exractlogfile, c-ip,cs method like GET and POST, cs-uri-stem, cs-uri-query,

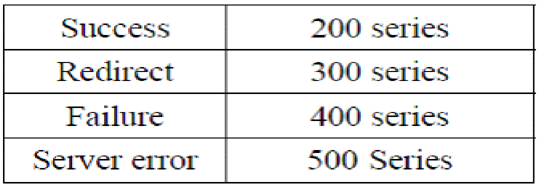
tatus and format them in desired style. Fig.3 gives field extraction logs.



logFie	timestamp	c-ip	cs-method	sc-status
u_ex110615.log	15-06-2011 02:10:25	49.15.200.181	GET	404
u_ex110615.log	15-06-2011 02:18:37	49.15.200.181	GET	404
u_ex110615.log	15-06-2011 02:26:47	49.15.200.181	GET	404
u_ex110615.log	15-06-2011 02:32:49	115.113.20.199	GET	200
u_ex110615.log	15-06-2011 02:32:55	115.113.20.199	GET	200
u_ex110615.log	15-06-2011 02:34:47	115.113.20.199	GET	200
u_ex110615.log	15-06-2011 03:37:34	203.90.64.100	GET	200
u_ex110615.log	15-06-2011 03:44:35	80.239.243.240	GET	200
u_ex110615.log	15-06-2011 03:45:27	80.239.243.240	GET	404
u_ex110615.log	15-06-2011 03:49:38	117.204.220.252	GET	200

Fig 3. Field Extraction Logs

The data cleaning is intended to clean web log data by deleting irrelevant and useless records [9]. A website can be accessed by millions of users so records with failed HTTP status code also may involve in the log data. Delete status code which is less than 200 and greater than 399. Classes of status code shown in fig.4. Remove logs entry nodes contain extension like .jpg, .gif, .ico, .img,.css means remove request such as multimedia files, image, page style file. Select only requested method GET. Therefore some of the entries are useless for analysis process, that is cleaned from log file.



Success	200 series
Redirect	300 series
Failure	400 series
Server error	500 Series

Fig 4. Classes of status code

Step 3: User identification and session Identification

User Identification, it is not important to identify the identity of the user, what is important is to detect attack he/she triggers if they're. To identify unique user we propose some rule. If there is new IP address then there is a new user. If an IP address is same but operating system or browsing software are different , a reasonable assumption is that each different agent type for an IP address represent a different user [10]. Data of records after cleaning is shown in Table 3. Statistical report 2 is about no. of records in GET method only status code 200 series and 300 series shown in Table 4.Session Identification, A user session can be defined as a set of pages visited by the same user within the duration of one particular visit to a website [11]. Session identification is time oriented. If time between requests exceeds 30 minutes, a new session is started. A user may have a single or multiple session during a period. When each user has been identified , an entry for each user must be divided into sessions. A timeout is used to break the entry into sessions. The Session timeout is set to 30 minutes during the experiment.

Step 4: Learning phase

We used Apriorialgorithm , which can show that attribute value frequently appear together in given log record, also it can mine relationship between attribute values from database table. Apriori algorithm is also known as association rule of mining algorithm. We take set of URIs with parameter list like cs-uri-stem, cs-uri-query and find frequent parameter list .Classification is to classify each audit record into one of the possible categories namely normal or a particular kind of intrusions[12]. By using Apriori algorithm find frequent rules having confidence greater. Following example rules finding after Apriori Algorithm.

"/main.asp",confidence=100%
"/login.asp" →"username" confidence=100%
"/login.asp" → "username", "password", confidence=75%

"/main.asp" → mpage=contactus.html,confidence=75%

Step 5: Detection Phase

Learning phase techniques, which build a model of normal behavior. Once the model of normal behavior is established. Then our IDS switches to, Detection phase which compare behavior of learning if abnormal behavior comes that is an intrusion. For each parameter list if forbidden keyword like select,insert,update,<script>,exec is found then that is SQL injection ans CSS attack. When unexpected event occur mean abnormal behavior that is anomaly attack. But we detect anomaly by using login threshold and frequency of page threshold.

Step 6: No. of attack shows type viz.

Following analyze attack detection is output of system. Attack Analysis shown in Table 5 .

We used 60% of the log data for training, and 40% of the log data to test for intrusion detection.The result of the two logs analyzed with IDS and the intrusions detected. The goal of this experiment is to prove that our IDS can detect more or same intrusion with other system. Our proposed IDS is efficient for detect intrusion in both misuse and anomaly mode by using web server access log file. There are a few system dedicated to web intrusion detection like Swatch, Logscaner and SNORT [7]. We compared our system with two other system SNORT and SensorWebIDS. SNORT which detect some web intrusion, The Sensor WebIDS is tested using log file of web server[7]. Detection Rate,It is the ratio between the total numbers of attack connections detected by our proposed model to the total number of attacks currently available in the data set. Detection rate formula is given below[13].

$$DetectionRate(DR) = \frac{Totaldetectedattacks}{TotalAttacks} * 100$$

SNORT having 33.3% detection rate for web server log file[7]. SensorWebIDS having 98.3% Detection rate[7]. Sensor WebIDS is good for detect misues and anomaly based attacks in web server log file. Comparision of detection rate graph shown in figure 5. Our system Detection rate is 98.5% . We calculate Detction rate to prove our efficiency, it is approximately same as SensorWebIDS.

5. EXPERIMENTAL RESULT

Our propose system have been implemented using visual studio 2008,.Net framework, using c# programming language under window7 operating system. Goal of an intrusion detection system is to detect that bad thing are happening, detect Previouslyknow and unknown attacks. Following are tables gives result on preprocessing web logs for webIDS.

Table 1. Results of data before cleaning

Log File	No. of Records	No. of Users	URLs
Log1	13409	311	2472
Log2	344	19	83

Table 2. Statistical report 1

Log File	.gif	.jpg	.ico	.png	.img	.css
Log1	5480	1299	378	1	377	308
Log2	186	26	20	17	10	22

Table 3. Results of data after cleaning

Log File	No. of Records	No. of Users	URLs
Log1	1980	250	130
Log2	57	16	9

Table 4.Statistical report 2

Log File	Status code	Method	Records
Log1	200	GET	1600
Log1	304	GET	64
Log1	206	GET	299
Log1	302	GET	5
Log1	301	GET	12
Log2	304	GET	6
Log2	200	GET	51

Table 5.Attack Analysis

Log File	No. of Records	SQL	XSS	Anomaly
Log1	1980	187	1	11
Log2	57	3	2	-

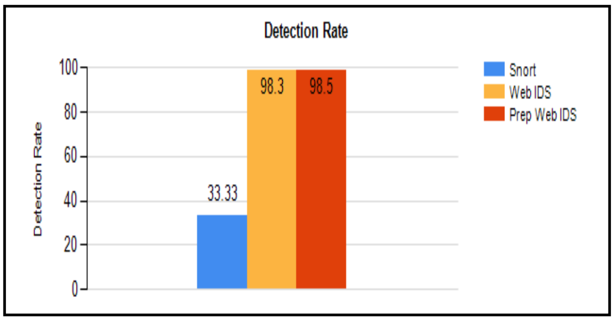


Fig 5. Result of detection rate

6. CONCLUSION

An intrusion detection system (IDS) is an increasingly important part of the security. In this Paper , we have presented on web intrusion detection mechanism using host based intrusion detection for both techniques misused based and anomaly based.This paper has presented details of preprocessing tasks that are necessary for performing intrusion detection. In Preprocessing , reduce log file size and increase quality of available data. Data cleaning, errors and inconsistencies are detected and removed to improve the quality of data. User Identification means identifying individual users by observing their IP address. Host based Intrusion Detection System monitor activity on a single host. Misused and anomaly both method used to detect intrusion or attack. Intrusion detection is a process of gathering intrusion related knowledge occurring during the system monitoring, and then analyzing collected data to draw aconclusion whether the system is intrusive or nor.

7. REFERENCES

[1] Wang Pu, Wang Jun-qing,2011,Intrusion Detection System with the Data Mining Technologies, pp. 490-492.

[2] ShaimaaSalama et.al. ,2011, Web server logs for preprocessing for web intrusion detection, Published by Canadian Center of Science and Education, Vol. 4, No. 4, pp. 123-133.

[3] L.K. Joshila Grace, V.Maheswari, DhinaharanNagamalai, 2011, Analysis of web logs and web user in web Mining, IJNSA, Vol.3, No.1, pp. 99-111.

[4] TheintTheint Aye, 2011, Web Log Cleaning for Mining Of Web Usage Patterns, IEEE, pp. 490- 494.

[5] Duanyang Zhao, QingxiangXu, Zhilin Feng,2010,Analysis and Design for Intrusion Detection System Based on Data Mining, pp. 339-342.

[6] Changxin Song, Ke Ma, 2009, Design of Intrusion Detection System Based on Data Mining Algorithm,ICSPS IEEE, pp. 370-373.

[7] C.J. Ezeife, J. Dong, A.K. Aggarwal, 2007, SensorWebIDS: A Web Mining Intrusion Detection System,International Journal of Web Information Systems, volume 4, pp. 97-120.

[8] Giovanni Vigna, William Robertson, Vishal Kher Richard, A. Kemmerer, 2003, A Stateful Intrusion Detection System forWorld-Wide Web Servers, ACSAC-IEEE,pp.1-10.

[9] G. Castellano, A. M. Fanelli, M. A. Torsello, 2007, Log Data Preparation for Mining Web Usage Patterns, IADIS International Conference Applied Computing, pp. 371-378.

[10] K.R. Suneetha, Dr. R. Krihnamoorthi, 2009, Identifying User Behavior by Analyzing Web Server Access Log File,IJCSNS , Vol. 9, No. 4, pp. 327-332.

[11] V. Chitraa, Dr. A.S. Davamani, 2010, A Survey on Preprocessing Methods for Web Usage Data, IJCSIS, Vol.7,No. 3, pp. 78-80..

[12] A Murali, M Rao,,2005, A Survey on Intrusion Detection Approaches,IEEE, pp. 233-244.

[13] Natesan. P, P. Balasubramanie , G. Gowrison, 2012, Improving the Attack Detection Rate in Network Intrusion Detection using Adaboost Algorithm, Journal of Computer Science,pp.1041-1048.