An Approach to Understand the End User Behavior through Log Analysis

Nikhil Kumar Singh
Department of
Computer Science and Engineering
Maulana Azad
National Institute of Technology
Bhopal, India

Deepak Singh Tomar
Department of
Computer Science and Engineering
Maulana Azad
National Institute of Technology
Bhopal, India

Bhola Nath Roy
Department of
Computer Science and Engineering
Maulana Azad
National Institute of Technology
Bhopal, India

ABSTRACT
Categorizing the end user in the web environment is a mind-numbing task. Huge amount of operational data is generated when end user interacts in web environment. This generated operational data is stored in various logs and may be useful source of capturing the end user activates. Pointing out the suspicious user in a web environment is a challenging task. To conduct efficient investigation in cyber space the available logs should be correlated. In this paper a prototype system is developed and implemented which is based on relational algebra to build the chain of evidence. The prototype system is used to preprocess the real generated data from logs and classify the suspicious user based on decision tree. At last various challenges in the logs managements are presented.

Keywords
cyber forensic; log file; correlation; decision tree,chain of evidence ,cyber crime;.

1. INTRODUCTION

Log files are like the black box on an aero plane that records the events occurred within an organization's system and networks. Logs are composed of log entries that play a very important role in evidence gathering and each entry contains information related to a specific event that has occurred within a system or a network. Log files helps cyber forensic process in probing and seizing computer, obtaining electronic evidence for criminal investigations and maintaining computer records for the federal rules of evidence.

Cyber forensic is a analytical method for extracting information and data from victim's computer storage media, it follows a systematic approach for finding, collecting, preserving data that guarantees information accuracy, reliability and presenting all evidence in acceptable manner in court of law. The primary goal of Cyber forensic is to reduce investigation time and complexity, it's not designed to solve crime but narrow the investigation. Cyber Forensics is done on cyber-crimes, password breaking, spamming, data recovery and analysis, tracking user activity, forensic imaging & verification, viruses, file types (extensions), encryption, Hacking etc.[1,2,3].

The establishment of cyber forensic is as same way as science and art is at its earlier stage. But the method of auditing, security, and law enforcement in cyber forensic changes at rapid pace as technology evolves with time. Even almost daily, new strategies, procedures and models are developed for forensic professionals in order to find electronic evidence, collecting it, preserving it, and presenting it in better way to use it potentially in the prosecution of cyber criminals [4].

Some logs are generally more likely than others to record information that may be useful in several situations, such as attack detection, fraud and misuse. For each type of situation, some records are generally more likely to contain detailed information about the activities in question. Other records typically contain less detailed information, and are often only useful to correlate the events recorded in the main log types. For example, an intrusion detection system can record the malicious commands issued to a server from an external host, this would be a primary source of attack information. A manager of an accident then might consider a firewall logs in search of other attempts to connect the source IP address, which is a secondary source of computer attacks [5].

The research described in this paper focuses on the nature of the event information provided in commonly available computer and other log and the extent to which it is possible to correlate such event information despite its heterogeneous nature and origins.

2. LOG FILES

Log files are excellent sources for determining the health status of a system and are used to capture the events happened within an organization's system and networks. Logs are a collection of log entries and each entry contains information related to a specific event that has taken place within a system or network. Many logs within an association contain records associated with computer security which are generated by many sources, including operating systems on servers, workstations, networking equipment and other security software's, such as antivirus software, firewalls, intrusion detection and prevention systems and many other applications. Routine log analysis is beneficial for identifying security incidents, policy violations, fraudulent activity, and operational problems. Logs are also useful for performing auditing and forensic analysis, supporting internal investigations, establishing baselines, and identifying operational trends and long-term problems.

Initially, logs were used for troubleshooting problems, but nowadays they are used for many functions within most organizations and associations, such as optimizing system and network performance, recording the actions of users, and providing data useful for investigating malicious activity. Logs have evolved to contain information related to many different types of events occurring within networks and systems. Within an organization, many logs contain records related to computer security; common examples of these computer security logs are

audit logs that track user authentication attempts and security device logs that record possible attacks.

With the world wide deployment of network servers, service station and other computing devices, the number of threats against networks and systems have greatly increased in number, volume, and variety of computer security logs and with the revolution of computer security logs, computer security log management are required. Log management is essential to ensure that computer security records are stored in sufficient detail for an appropriate period of time. Log management is the process for generating, transmitting, storing, analyzing, and disposing of computer security log data. The fundamental problem with log management is effectively balancing a limited quantity of log management resources with a continuous supply of log data. Log generation and storage can be complicated by several factors, including a high number of log sources; inconsistent log content, formats, and timestamps among sources; and increasingly large volumes of log data [5,6]. Log management also involves protecting the confidentiality, integrity, and availability of logs. Another problem with log management is ensuring that security, system, and network administrators regularly perform effective analysis of log data.

3. TROUBLES IN LOG MANAGEMENT

In an association, many Operating Systems, security software, and other applications generate and preserve their independent log files. This complicates log management in the following ways [5, 7]

3.1 Multiple Log Sources

Logs can be found on many hosts throughout the organization that should be required to conduct log management throughout the organization. In addition, a single log source can generate multiple logs for example, an application storing authentication attempts in one log and network activity in another log

3.2 Heterogeneous Log Content

Log file capture certain pieces of information in each entry, such as client and server IP addresses, ports, date and time etc. For efficiency, log sources often record only the pieces of information that they consider most important. It creates difficulty to make an relationship between event records and different log sources because they may not have any common attribute (e.g., source 1 records the source IP address but not the username, and source 2 records the username but not the source IP address). Even the representation of log value varies with log source; these differences may be slight, such as one date being in YYYYMMDD format and another being in MMDDYYYY format, or they may be much more complex.

3.3 Inconsistent Timestamps

Usually every application who ganerates logs uses the local timestamps i.e. the timestamps of the internal clock. If the host's clock is not synchronized or inaccurate, then log file analysis is more difficult, specially when the environment has multiple hosts. For example, timestamps may indicate that event "X" happened 2 minutes after event "Y", whereas event 'X' has actually happened 55 seconds before event "Y".

3.4 Multiple Log Formats

Each application that creates logs may use its own format, eg. XML format or SNMP format, comma-separated or tab-separated format, and other binary formats. Some logs are designed in a way that they are readable to humans, whereas some others don't, some use standard formats, whereas others use proprietary formats. Some logs are such that they are not stored on a single host, but are transmitted to other hosts for processing; a common example can be SNMP traps.

3.5 Log Confidentiality and Integrity

Protection of Log records to maintain their integrity and confidentiality is very essential and challenging. For example, logs might intentionally or unintentionally capture sensitive information such as users' passwords and the content of e-mails. This raises security and privacy concerns relating both the individuals that examine the logs and others that might be able to access the logs through authorized or unauthorized means. Logs which are secured improperly in storage or in transit might also be susceptible to intentional and inadvertently alteration and destruction. This could cause a variety of impacts, including allowing malicious activities to go unnoticed and manipulating evidence to conceal the identity of a malicious party.

Protection of logs availability is also a very big issue. Many logs having a size limit when this limit is reached, the log might overwrite old data with new data or stop logging all together both of which would cause a loss of log data availability. To meet data retention requirements, it's necessary to establish log archival i.e. keeping copies of log files for a longer period of time than the original log sources can support. Because of the volume of logs, it might be appropriate in some cases to reduce the logs by filtering out log entries that do not need to be archived. The confidentiality and integrity of the archived logs also need to be protected

4. ROLE OF EVENT LOG DATA IN EVIDENCE GATHERING

Logs are composed of log entries; each entry contains information related to a specific event that has occurred within a system or network. If the suspicious end user exploits web form as an access point for input attacks like cross-site scripting, SQL injection and buffer overflow attack on a web application, it may be detected using the log file [5].An interesting question is raised, why event data should be logged on a given system. Essentially there are four categories of reasons.

4.1 Accountability

Log file data can be used to identify which type of accounts are associated with certain events and that information can be used to emphasize where training and/or disciplinary actions are needed.

4.2 Rebuilding

What was happening before and during an event can be reviewed chronologically by using log file data. For this it should be ensured that the clocks are regularly synchronized to a central source to ensure that the date/time stamps are in synchronization.

4.3 Intrusion Detection

Log data can be reviewed for detecting unusual or unauthorized events, assuming that the correct data is being logged and reviewed. But variation of unusual activities is a main problem i.e. login attempts outside of designated schedules, failed login attempts, port sweeps, locked accounts, network activity levels, memory utilization, key file/data access, etc.

4.4 Problem Recognition

Log data can be used for problem recognition and to identify security events, for example resource utilization, investigating causal factors of failed and so on.

5. PROPOSED FRAMEWORK

The Details of Proposed framework are as follows:

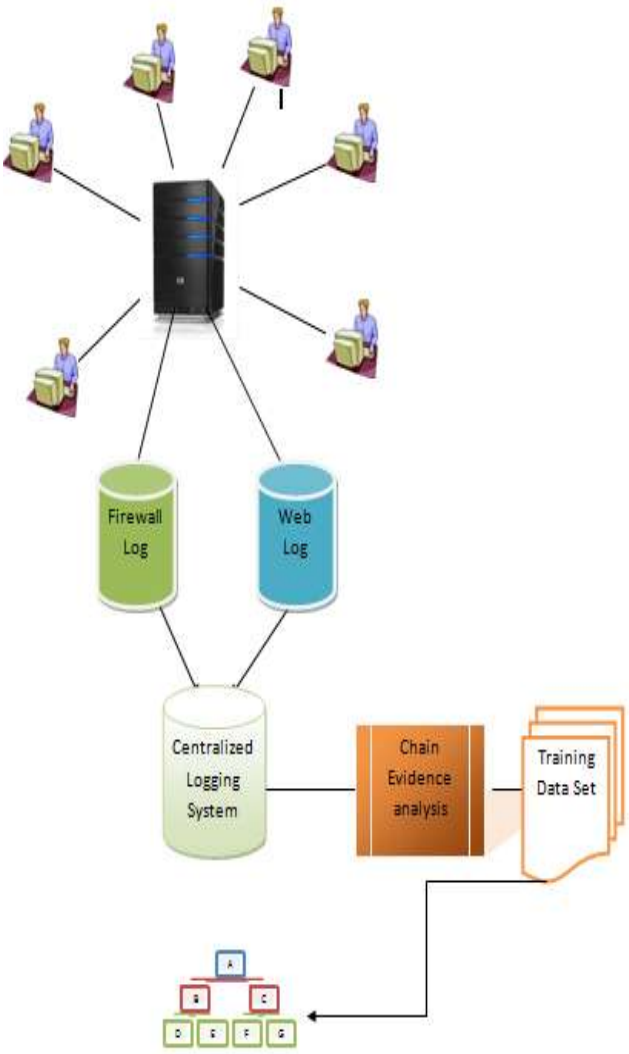


Fig. 1 Proposed Framework

5.1 Centralization of Log files

In this step, log files maintained by the web server and firewall are extracted and stored in the central location. The data are transformed in a suitable format for conducting effective analysis..

5.2 Chain of evidence analyzer

The evidence analyzer takes the firewall log and web log from the centralized log. It applies the rule based correlation by URLs and Time techniques as shown in Section VI and creates the training data set.

5.3 Decision Tree Construction

In this step a decision tree is constructed from the resultant training data set by applying decision tree algorithm.

6. METHODOLOGY

Individual log files records activities related to a particular application although useful in many developments contain a lot of data that might not be particularly useful in evidence gathering. However, comparative analysis of different types of log files, coming from different applications run on the same host or different host can reveal useful interrelations that can be used in evidence gathering. This work deals with comparative analysis of firewall log file and web server log file. Firewalls log event fall into three broad categories: critical system issues (hardware failures and the like), significant authorized administrative events (rule set changes, administrator account changes), and network connection logs. Interesting information present in the firewall log required for our proposed framework can be, changes to firewall policy, addition/deletion/change in administrative accounts, network connection logs of the compromised system, which include dropped and rejected connections, time/protocol/IP addresses/usernames for allowed connections, amount of data transferred etc. Web server log 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, the accessed page etc. This information is very useful.

This work is concentrated on correlations of firewall logs and web logs coming from different applications running on the same host, as well as correlations of logs coming from different (or the same) applications running on different hosts during the same period of time, a decision tree is also developed on the basis of that correlated information, helps in taking proper decision. During the initial pre-processing firewall log and web log from web server is accessed and the client's ip address in firewall logs which probes cross the threshold limit within a fix time period is determined, if this ip address also present in web log and it accesses the restricted area then it's suspicious for server.

7. MODEL FOR WORK

7.1 Algebraic Representation:

Symbol	**Meaning**
l_1	Firewall Log
l_2	Web Log
f_{Ip}	Set of client IP address of firewall log
f_{Dp}	Set of destination port of firewall log

n_{Ip}	Normal Users IP address
s_{Ip}	Suspicious user IP address
sw_{Ip}	Suspicious web user IP address
at_{Ip}	Attacker IP address

$$f_{Ip} = \{Ip | Ip\ is\ a\ IP\ Address\ of\ client\ in\ l_1\}$$
$$f_{Dp} = \{Dp | Dp\ is\ the\ Destination\ Port\ in\ l_1\}$$
$$w_{Ip} = \{Ip | Ip\ is\ a\ IP\ Address\ of\ client\ in\ l_2\}$$
$$R(f_{Ip}, f_{Dp}) \rightarrow Relation\ of\ f_{Ip}\ and\ f_{Dp}$$

** There is many to many relationship between f_{Ip} and f_{Dp}*

$s_{Ip} =$
$\{Ip | Ip\ having\ more\ than\ 200\ probes(image)\ in\ relation\ R\}$

$r_{Ip} =$
$\{Ip | Ip\ is\ a\ IP\ Address\ of\ client\ having\ restricted\ zone\ entry\ in\ l_2\}$

$$n_{Ip} = (f_{Ip} - s_{Ip})$$

$$sw_{Ip} = (s_{Ip} - w_{Ip}) + (s_{Ip} \cap w_{Ip})$$
$$at_{Ip} = (sw_{Ip} - r_{Ip}) + (sw_{Ip} \cap r_{Ip})$$

7.2 Relational Algebra Representation:

Symbol	**Meaning**
L_1	Firewall Log
L_2	Web Log
$\mathcal{G}$	Aggregate Function

$$temp_1 \longleftarrow f_{cip}\mathcal{G}_{\text{count\_distinct(dest\_port)}}(L_1)$$
$$temp_2 \longleftarrow temp_1 \bowtie \rho_{L_3(w_{cip})}(L_2)$$
$$temp_3 \longleftarrow \prod_{w_{cip}} (\sigma_{webentry=\text{"Admin" } OR \text{ "Set"}}(L_2))$$
$$temp_4 \longleftarrow temp_2 \bowtie temp_3$$
$$\prod_{f_{cip}} (\sigma_{probes < 200}(temp_1)) = normal\ user$$
$$\prod_{f_{cip}} (\sigma_{probes > 200}(temp_1)) = suspicious\ user$$
$$\prod_{f_{cip}} (\sigma_{probes > 200,}(temp_3)) = suspicious\ user\ for\ web$$
$$\prod_{f_{cip}} (\sigma_{probes > 200,}(temp_4)) = attacker$$

7.3 Decision Tree

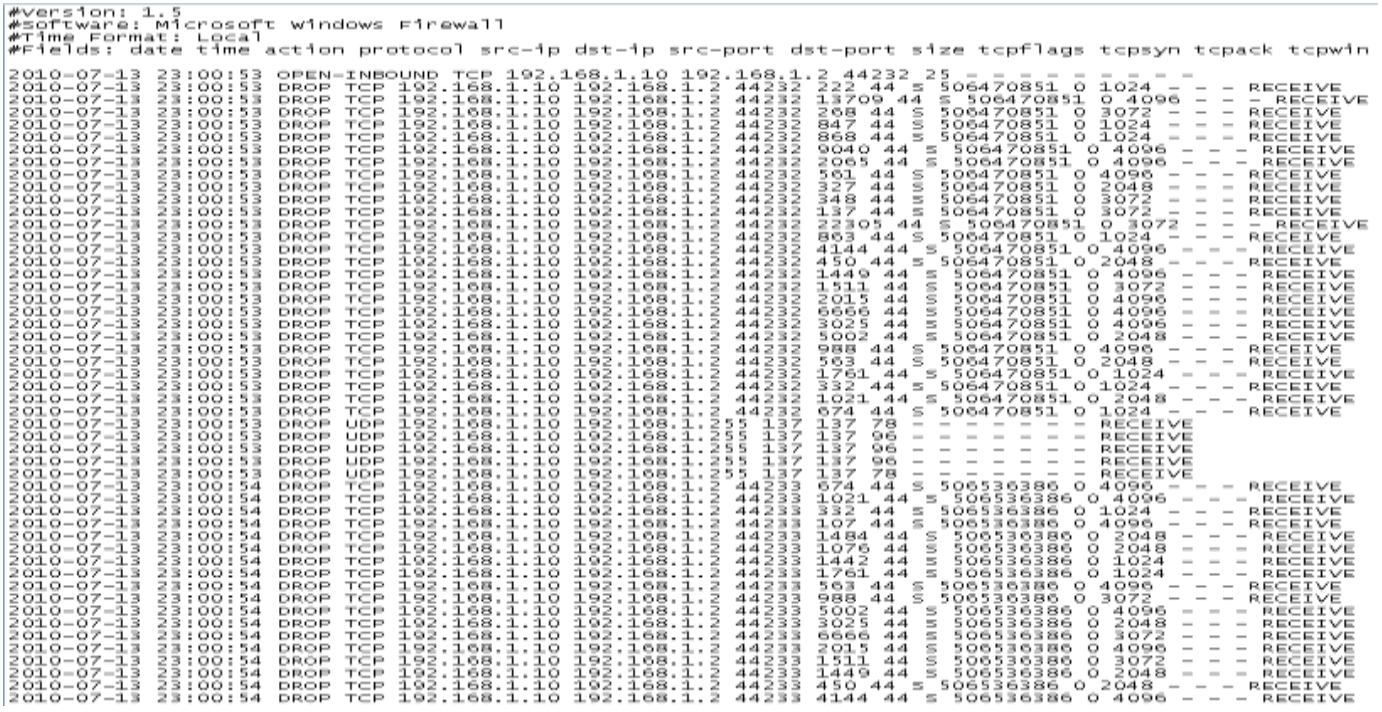
A decision tree is a tool which is used to support decisions, this tool uses a tree-like graph or model of decisions and their possible consequences, Decision tree are generally used where there is a need of decision analysis such as operations research, this helps in identifying a strategy which will help in reaching a goal. Decision tree is also used to calculate conditional probabilities.

7.4 Decision tree induction

A decision tree is flow charts like tree structure, where each internal node denotes a test or decision on an attribute, each branch represents an output come of the test and leaf node represent class distributions. Selection of test attribute at each node in the tree is based on information gain of that attribute, a attributes having highest information gain selected as test attributes. To calculate information gain of any attribute, assume a set D having *d* data sample, suppose the class level attribute having *n* distinct values defined as *n* distinct classes C_i(for i=1,……. n), then the expected information needs to be classified to a given sample.

$$I(d_1, d_2, \ldots\ldots\ldots\ldots, d_n) = -\sum_{i=1}^{n} p_i log_2(p_i)$$

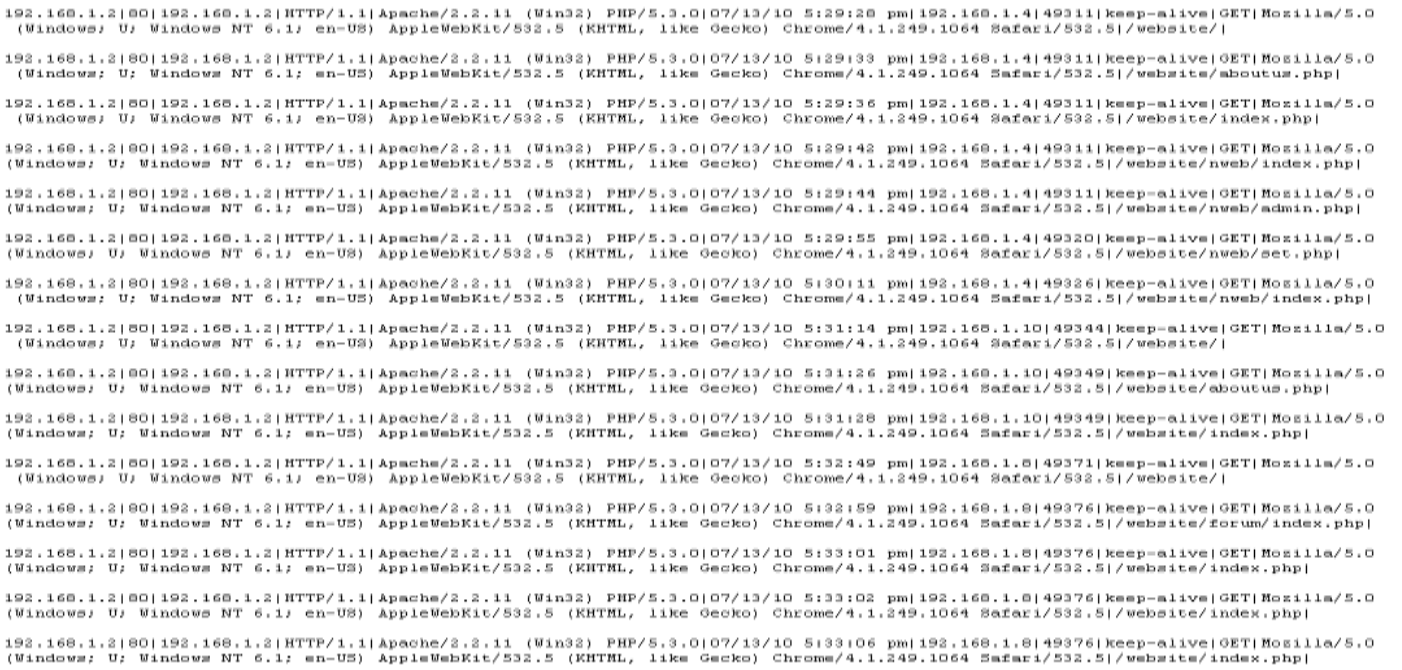
Where p_i is the probability of any attribute sample belongs to the class C_i. Let us assume attribute *X* having *m* distinct values, {x_1, x_2, x_3,…..,x_m}. Attribute X can be used to partition *D* into *m* subsets, {D_1, D_2, D_3,……, D_m} where D_j contains those



```
#Version: 1.5
#Software: Microsoft Windows Firewall
#Time Format: Local
#Fields: date time action protocol src-ip dst-ip src-port dst-port size tcpflags tcpsyn tcpack tcpwin

2010-07-13 23:00:53 OPEN-INBOUND TCP 192.168.1.10 192.168.1.2 44232 25 - - - - - - - - -
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 222 44 - S 506470851 0 1024 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 13709 44 - S 506470851 0 4096 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 268 44 - S 506470851 0 3072 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 847 44 - S 506470851 0 1024 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 868 44 - S 506470851 0 1024 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 9040 44 - S 506470851 0 4096 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 2065 44 - S 506470851 0 4096 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 561 44 - S 506470851 0 4096 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 327 44 - S 506470851 0 2048 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 348 44 - S 506470851 0 3072 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 137 44 - S 506470851 0 3072 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 22305 44 - S 506470851 0 3072 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 863 44 - S 506470851 0 1024 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 4144 44 - S 506470851 0 4096 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 450 44 - S 506470851 0 2048 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 1449 44 - S 506470851 0 4096 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 1511 44 - S 506470851 0 3072 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 2015 44 - S 506470851 0 4096 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 6666 44 - S 506470851 0 4096 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 3025 44 - S 506470851 0 4096 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 5002 44 - S 506470851 0 2048 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 988 44 - S 506470851 0 4096 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 563 44 - S 506470851 0 2048 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 1761 44 - S 506470851 0 1024 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 332 44 - S 506470851 0 1024 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 1021 44 - S 506470851 0 2048 - - - RECEIVE
2010-07-13 23:00:53 DROP TCP 192.168.1.10 192.168.1.2 44232 674 44 - S 506470851 0 1024 - - - RECEIVE
2010-07-13 23:00:53 DROP UDP 192.168.1.10 192.168.1.255 137 137 78 - - - - - - - - RECEIVE
2010-07-13 23:00:53 DROP UDP 192.168.1.10 192.168.1.255 137 137 96 - - - - - - - - RECEIVE
2010-07-13 23:00:53 DROP UDP 192.168.1.10 192.168.1.255 137 137 96 - - - - - - - - RECEIVE
2010-07-13 23:00:53 DROP UDP 192.168.1.10 192.168.1.255 137 137 96 - - - - - - - - RECEIVE
2010-07-13 23:00:53 DROP UDP 192.168.1.10 192.168.1.255 137 137 78 - - - - - - - - RECEIVE
2010-07-13 23:00:54 DROP TCP 192.168.1.10 192.168.1.2 44233 674 44 - S 506536386 0 4096 - - - RECEIVE
2010-07-13 23:00:54 DROP TCP 192.168.1.10 192.168.1.2 44233 1021 44 - S 506536386 0 4096 - - - RECEIVE
2010-07-13 23:00:54 DROP TCP 192.168.1.10 192.168.1.2 44233 332 44 - S 506536386 0 1024 - - - RECEIVE
2010-07-13 23:00:54 DROP TCP 192.168.1.10 192.168.1.2 44233 107 44 - S 506536386 0 4096 - - - RECEIVE
2010-07-13 23:00:54 DROP TCP 192.168.1.10 192.168.1.2 44233 1484 44 - S 506536386 0 2048 - - - RECEIVE
2010-07-13 23:00:54 DROP TCP 192.168.1.10 192.168.1.2 44233 1076 44 - S 506536386 0 2048 - - - RECEIVE
2010-07-13 23:00:54 DROP TCP 192.168.1.10 192.168.1.2 44233 1442 44 - S 506536386 0 1024 - - - RECEIVE
2010-07-13 23:00:54 DROP TCP 192.168.1.10 192.168.1.2 44233 1761 44 - S 506536386 0 1024 - - - RECEIVE
2010-07-13 23:00:54 DROP TCP 192.168.1.10 192.168.1.2 44233 563 44 - S 506536386 0 4096 - - - RECEIVE
2010-07-13 23:00:54 DROP TCP 192.168.1.10 192.168.1.2 44233 988 44 - S 506536386 0 3072 - - - RECEIVE
2010-07-13 23:00:54 DROP TCP 192.168.1.10 192.168.1.2 44233 5002 44 - S 506536386 0 4096 - - - RECEIVE
2010-07-13 23:00:54 DROP TCP 192.168.1.10 192.168.1.2 44233 3025 44 - S 506536386 0 2048 - - - RECEIVE
2010-07-13 23:00:54 DROP TCP 192.168.1.10 192.168.1.2 44233 6666 44 - S 506536386 0 3072 - - - RECEIVE
2010-07-13 23:00:54 DROP TCP 192.168.1.10 192.168.1.2 44233 2015 44 - S 506536386 0 4096 - - - RECEIVE
2010-07-13 23:00:54 DROP TCP 192.168.1.10 192.168.1.2 44233 1511 44 - S 506536386 0 3072 - - - RECEIVE
2010-07-13 23:00:54 DROP TCP 192.168.1.10 192.168.1.2 44233 1449 44 - S 506536386 0 2048 - - - RECEIVE
2010-07-13 23:00:54 DROP TCP 192.168.1.10 192.168.1.2 44233 450 44 - S 506536386 0 2048 - - - RECEIVE
2010-07-13 23:00:54 DROP TCP 192.168.1.10 192.168.1.2 44233 4144 44 - S 506536386 0 4096 - - - RECEIVE
```

Fig. 2 End User probing in Sample Windows Firewall log



```
192.168.1.2|80|192.168.1.2|HTTP/1.1|Apache/2.2.11 (Win32) PHP/5.3.0|07/13/10 5:29:28 pm|192.168.1.4|49311|keep-alive|GET|Mozilla/5.0
(Windows; U; Windows NT 6.1; en-US) AppleWebKit/532.5 (KHTML, like Gecko) Chrome/4.1.249.1064 Safari/532.5|/website/|

192.168.1.2|80|192.168.1.2|HTTP/1.1|Apache/2.2.11 (Win32) PHP/5.3.0|07/13/10 5:29:33 pm|192.168.1.4|49311|keep-alive|GET|Mozilla/5.0
(Windows; U; Windows NT 6.1; en-US) AppleWebKit/532.5 (KHTML, like Gecko) Chrome/4.1.249.1064 Safari/532.5|/website/aboutus.php|

192.168.1.2|80|192.168.1.2|HTTP/1.1|Apache/2.2.11 (Win32) PHP/5.3.0|07/13/10 5:29:36 pm|192.168.1.4|49311|keep-alive|GET|Mozilla/5.0
(Windows; U; Windows NT 6.1; en-US) AppleWebKit/532.5 (KHTML, like Gecko) Chrome/4.1.249.1064 Safari/532.5|/website/index.php|

192.168.1.2|80|192.168.1.2|HTTP/1.1|Apache/2.2.11 (Win32) PHP/5.3.0|07/13/10 5:29:42 pm|192.168.1.4|49311|keep-alive|GET|Mozilla/5.0
(Windows; U; Windows NT 6.1; en-US) AppleWebKit/532.5 (KHTML, like Gecko) Chrome/4.1.249.1064 Safari/532.5|/website/nweb/index.php|

192.168.1.2|80|192.168.1.2|HTTP/1.1|Apache/2.2.11 (Win32) PHP/5.3.0|07/13/10 5:29:44 pm|192.168.1.4|49311|keep-alive|GET|Mozilla/5.0
(Windows; U; Windows NT 6.1; en-US) AppleWebKit/532.5 (KHTML, like Gecko) Chrome/4.1.249.1064 Safari/532.5|/website/nweb/admin.php|

192.168.1.2|80|192.168.1.2|HTTP/1.1|Apache/2.2.11 (Win32) PHP/5.3.0|07/13/10 5:29:55 pm|192.168.1.4|49320|keep-alive|GET|Mozilla/5.0
(Windows; U; Windows NT 6.1; en-US) AppleWebKit/532.5 (KHTML, like Gecko) Chrome/4.1.249.1064 Safari/532.5|/website/nweb/set.php|

192.168.1.2|80|192.168.1.2|HTTP/1.1|Apache/2.2.11 (Win32) PHP/5.3.0|07/13/10 5:30:11 pm|192.168.1.4|49326|keep-alive|GET|Mozilla/5.0
(Windows; U; Windows NT 6.1; en-US) AppleWebKit/532.5 (KHTML, like Gecko) Chrome/4.1.249.1064 Safari/532.5|/website/nweb/index.php|

192.168.1.2|80|192.168.1.2|HTTP/1.1|Apache/2.2.11 (Win32) PHP/5.3.0|07/13/10 5:31:14 pm|192.168.1.10|49344|keep-alive|GET|Mozilla/5.0
(Windows; U; Windows NT 6.1; en-US) AppleWebKit/532.5 (KHTML, like Gecko) Chrome/4.1.249.1064 Safari/532.5|/website/|

192.168.1.2|80|192.168.1.2|HTTP/1.1|Apache/2.2.11 (Win32) PHP/5.3.0|07/13/10 5:31:26 pm|192.168.1.10|49349|keep-alive|GET|Mozilla/5.0
(Windows; U; Windows NT 6.1; en-US) AppleWebKit/532.5 (KHTML, like Gecko) Chrome/4.1.249.1064 Safari/532.5|/website/aboutus.php|

192.168.1.2|80|192.168.1.2|HTTP/1.1|Apache/2.2.11 (Win32) PHP/5.3.0|07/13/10 5:31:28 pm|192.168.1.10|49349|keep-alive|GET|Mozilla/5.0
(Windows; U; Windows NT 6.1; en-US) AppleWebKit/532.5 (KHTML, like Gecko) Chrome/4.1.249.1064 Safari/532.5|/website/index.php|

192.168.1.2|80|192.168.1.2|HTTP/1.1|Apache/2.2.11 (Win32) PHP/5.3.0|07/13/10 5:32:49 pm|192.168.1.8|49371|keep-alive|GET|Mozilla/5.0
(Windows; U; Windows NT 6.1; en-US) AppleWebKit/532.5 (KHTML, like Gecko) Chrome/4.1.249.1064 Safari/532.5|/website/|

192.168.1.2|80|192.168.1.2|HTTP/1.1|Apache/2.2.11 (Win32) PHP/5.3.0|07/13/10 5:32:59 pm|192.168.1.8|49376|keep-alive|GET|Mozilla/5.0
(Windows; U; Windows NT 6.1; en-US) AppleWebKit/532.5 (KHTML, like Gecko) Chrome/4.1.249.1064 Safari/532.5|/website/forum/index.php|

192.168.1.2|80|192.168.1.2|HTTP/1.1|Apache/2.2.11 (Win32) PHP/5.3.0|07/13/10 5:33:01 pm|192.168.1.8|49376|keep-alive|GET|Mozilla/5.0
(Windows; U; Windows NT 6.1; en-US) AppleWebKit/532.5 (KHTML, like Gecko) Chrome/4.1.249.1064 Safari/532.5|/website/index.php|

192.168.1.2|80|192.168.1.2|HTTP/1.1|Apache/2.2.11 (Win32) PHP/5.3.0|07/13/10 5:33:02 pm|192.168.1.8|49376|keep-alive|GET|Mozilla/5.0
(Windows; U; Windows NT 6.1; en-US) AppleWebKit/532.5 (KHTML, like Gecko) Chrome/4.1.249.1064 Safari/532.5|/website/index.php|

192.168.1.2|80|192.168.1.2|HTTP/1.1|Apache/2.2.11 (Win32) PHP/5.3.0|07/13/10 5:33:06 pm|192.168.1.8|49376|keep-alive|GET|Mozilla/5.0
(Windows; U; Windows NT 6.1; en-US) AppleWebKit/532.5 (KHTML, like Gecko) Chrome/4.1.249.1064 Safari/532.5|/website/index.php|
```

Fig. 3 Sample Web Server Log

samples in *D* that have a value x_j of A. The entropy based on the partition into the subset X is given by

$$E(X) = \sum_{j=1}^{m} \frac{d_{1j} + \ldots\ldots\ldots + d_{nj}}{d} I(d_{1j}, \ldots\ldots\ldots, d_{nj})$$

Where $\frac{d_{1j} + \ldots\ldots\ldots + d_{nj}}{d}$ is a ratio of number of sample in the subset to the total number of samples in D. Entropy value and the purity of subset partition is inversely proportional. The information gain is then calculated by

$$Gain(X) = I(d_1, d_2, \ldots\ldots\ldots\ldots, d_n) - E(X)$$

Whole these procedures compute the information gain of each and every attributes and attribute having highest information gain selected for test attributes for given set D.

8. IMPLEMENTATION DETAILS

A tool has been created and implemented, which takes firewall log and the web log of same time as input and generates a decision regarding client behavior whether client is normal user, suspicious user, suspicious web user or an attacker.

There were 70563 entries in firewall log of 30 clients that have accesses the server therefore determine the probes details of these 30 clients and calculate number of different ports probe by these 30 clients then take left outer join of these 30 firewall client ip address with client ip address of web log then again take left outer join of this resultant set with ip address of client having restricted area entry in web log file. The final resultant set describe as client (NU-normal user, SU-suspicious user, SUW- suspicious user for web, AT- attacker).

9. CONSTRUCTION OF DECISION TREE

This training data set (D) consists of 4 data samples (d=4) and 4 attributes. Initially use equation to compute the expected information need to classify these 4 samples.

$$I(d_1, d_2, d_3, d_4) = I(20,4,3,3) = -\frac{20}{30}\log_2\frac{20}{30} - \frac{4}{30}\log_2\frac{4}{30} - \frac{3}{30}\log_2\frac{3}{30} - \frac{3}{30}\log_2\frac{3}{30} = 1.44$$

Next we need to compute information gain of each attributes, let's start with attributes probes, probes having two different classes (yes, no).

For probes = "Yes"

$$I(d_{11}, d_{12}, d_{13}, d_{14}) = I(0,4,3,3) = 1.57$$

For probes = "no"

$$I(d_{21}, d_{22}, d_{23}, d_{24}) = 0$$

Above two equation show the expected information needed to classify probes then Entropy of probes is

IP	Probes	Web Entry	Restricted Zone	Decision
192.168.1.10	>=200	Yes	NO	SUW
192.168.1.12	>=200	Yes	Yes	AT
192.168.1.153	<=200	Yes	NO	NU
192.168.1.154	<=200	Yes	NO	NU
192.168.1.155	<=200	Yes	NO	NU
192.168.1.20	>=200	No	NO	SU
192.168.1.22	<=200	Yes	NO	NU
192.168.1.24	>=200	Yes	Yes	AT
192.168.1.26	<=200	Yes	NO	NU
192.168.1.28	<=200	Yes	NO	NU
192.168.1.33	<=200	Yes	NO	NU
192.168.1.35	<=200	Yes	NO	NU
192.168.1.37	<=200	Yes	NO	NU
192.168.1.4	>=200	Yes	Yes	AT
192.168.1.40	<=200	Yes	NO	NU
192.168.1.45	<=200	Yes	NO	NU
192.168.1.53	<=200	Yes	NO	NU
192.168.1.54	<=200	Yes	NO	NU
192.168.1.6	>=200	No	NO	SU
192.168.1.66	>=200	No	NO	SU
192.168.1.63	<=200	Yes	NO	NU
192.168.1.70	<=200	Yes	NO	NU
192.168.1.77	>=200	No	NO	SU
192.168.1.8	>=200	Yes	NO	SUW
192.168.1.88	>=200	Yes	NO	SUW
192.168.1.93	<=200	Yes	NO	NU
192.168.1.94	<=200	Yes	NO	NU
192.168.1.95	<=200	Yes	NO	NU
192.168.1.96	<=200	Yes	NO	NU
192.168.1.97	<=200	Yes	NO	NU

Figure 4 Sample Training Data Set Generated by Integrating Web Log & Firewall

Resultant set is an training data set contains of 5 attributes (Ip, probes, web_entry, restricted_zone, decision),where Ip describe the distinct client ip address of firewall log , probes describes describe where respective ip cross the probes threshold or not ,web_entry describe where respective ip having web entry or not , restricted zone entries describe where respective ip having restricted web entry or not and decision where show behavior of

$$E(probes) = \frac{10}{30} I(d_{11}, d_{12}, d_{13}, d_{14}) + \frac{20}{30} I(d_{21}, d_{22}, d_{23}, d_{24}) = .5233$$

Hence, the information gain

$$Gain(probes) = I(d_1, d_2, d_3, d_4) - E(probes) = .916$$

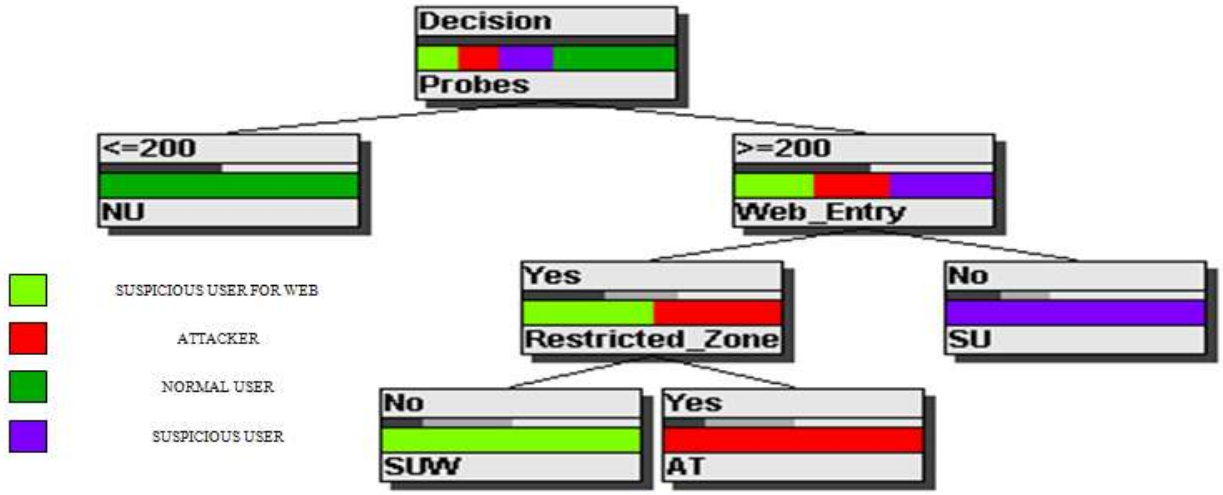


Fig. 5 Decision Tree

File Edit Format View Help

IP	Probes	Web Entry	Restricted Zone
192.168.1.123	>=200	Yes	No
192.168.1.124	>=200	Yes	Yes
192.168.1.201	>=200	No	No
192.168.1.224	<=200	Yes	No
192.168.1.241	>=200	Yes	Yes
192.168.1.209	<=200	Yes	No
192.168.1.204	<=200	Yes	No
192.168.1.207	<=200	Yes	No
192.168.1.135	<=200	Yes	No
192.168.1.137	<=200	Yes	No
192.168.1.104	>=200	Yes	Yes
192.168.1.140	<=200	Yes	No
192.168.1.145	<=200	Yes	No
192.168.1.156	<=200	Yes	No
192.168.1.166	>=200	No	No
192.168.1.163	<=200	Yes	No
192.168.1.170	<=200	Yes	No
192.168.1.177	>=200	No	No
192.168.1.189	<=200	Yes	No
192.168.1.188	>=200	Yes	No
192.168.1.120	<=200	Yes	No
192.168.1.231	>=200	No	No
192.168.1.146	<=200	Yes	No
192.168.1.134	<=200	Yes	No
192.168.1.139	<=200	Yes	No
192.168.1.178	<=200	Yes	No
192.168.1.167	<=200	Yes	No
192.168.1.108	>=200	Yes	No
192.168.1.112	<=200	YES	No
192.168.1.190	<=200	Yes	No
192.168.1.199	<=200	Yes	No

Figure 6 Sample Trained Data Set

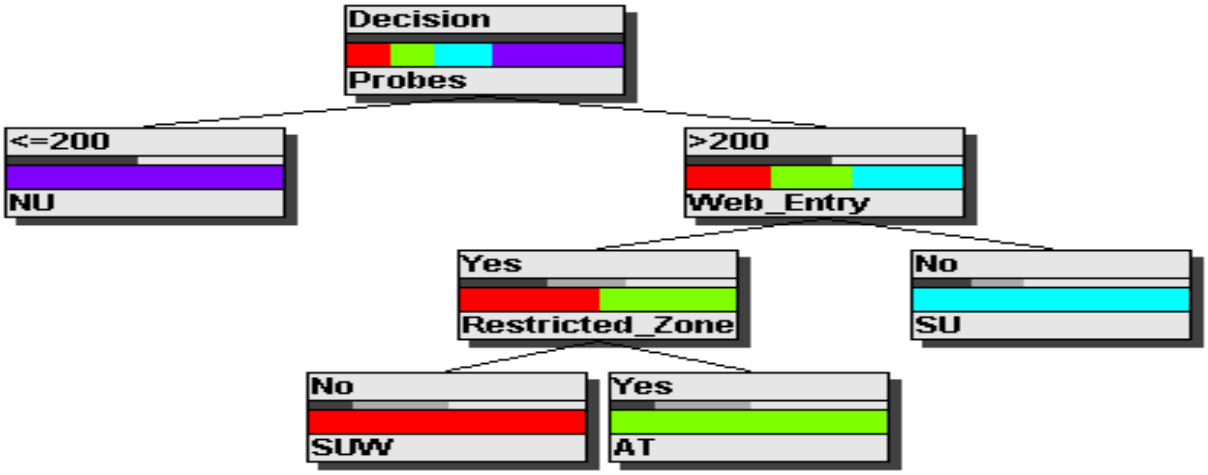


Figure 7 Decision tree generated from Trained Data Set

Similarly the Gain(web_entry) =0.563, Gain (restrict _zone) =0.4671 .Since probes has highest information gain among all the attributes so its selected as the test attributes. Decision tree is shown in Fig 5, describe the client behavior. To check the validity of decision tree take a trained data set and execute the generated rules again on the trained data set resulting in decision tree shown in Figure 7. It shows that the generated rules and decision tree stay correct for all cases.

9.1 Extracting Classification rules based on decision tree:

The decision tree of fig. 5 can be converted into classification IF-THEN rule by tracing the path from root node to each leaf node in the tree.

IF probes >=200, web_entry = "yes" and restricted_zone = "yes" then decision = "Attacker"

IF probes >=200, web_entry = "yes" and restricted_zone = "NO" then decision = "Suspicious User for web"

IF probes >=200, web_entry = "NO" and restricted_zone = "NO" then decision = "Suspicious User"

IF probes <=200, web_entry = "yes" and restricted_zone = "No" then decision = "Normal User"

10. CONCLUSIONS

In this work the implemented system extracts the evidence from different sources, relates generated logs on the basis of relational algebra and classifies suspicious user based on decision tree. The implemented system encourages the web administrator to study the navigation behavior of suspicious user and assist to enforce the effective security policy.

The future work will cover the issues related to log consistency, log integrity and log rotation

11. ACKNOWLEDGMENTS

The research presented in this paper would not have been possible without our college, at MANIT, Bhopal. We wish to express our gratitude to all the people who helped turn the World-Wide Web into the useful and popular distributed hypertext. We also wish to thank the anonymous reviewers for their valuable suggestions

12. REFERENCES

[1] http://www.all-about-forensic-science.com/cyber-forensics.html

[2] Gary L Palmer A Road Map for Digital Forensic Research. Technical ReportDTR-T0010-01, DFRWS. Report for the First Digital Forensic Research Workshop (DFRWS),(2001).

[3] Tamas Abraham "Event Sequence Mining to Develop Profiles for Computer Forensic Investigation Purposes" Information Networks Division Defence Science and Technology Organization, Australia

[4] http://www.cyberforensics.com

[5] Robert Rinnan "*Benefits of Centralized Log file Correlation*" Master's Thesis, Master of Science in Information Security30 ECTS, Department of Computer Science and Media Technology Gjøvik University College, 2005.

[6] Deepak Singh Tomar, J.L.Rana and S.C.Shrivastava, Evidence Gathering System for Input Attacks in (IJCNS) International Journal of Computer and Network Security Vol. 1, No. 1, October 2009.

[7] Muhammad Kamran Ahmed, Mukhtar Hussain and Asad Raza "*An Automated User Transparent Approach to log Web URLs for Forensic Analysis*" Fifth International Conference on IT Security Incident Management and IT Forensics 2009.

[8] Pavel Gladyshev "*Formalising Event Reconstruction in Digital Investigations*" Ph.D. dissertation Department of Computer Science, University College Dublin, 2004.

[9] Nabil HAMMOUD "Decentralized Log Event Correlation Architecture " MEDES, Lyon, France,2009

[10] Tamas Abraham and Olivier de Vel "*Investigative Profilling with Computer Forensic Log Data and Association*"IEEE,2002

[11] Data Mining – Concept and Techniques by Jiawei Han and Micheline Kamber.

Nikhil Kumar Singh M.Tech(Final Year) in Computer Science & Engg., B.E. in Computer Science and research scholar of Maulana Azad National Institute of Technology(MANIT), Bhopal.

Mr. Deepak Singh Tomar M.Tech & B.E. in Computer Science & Engg. and working as Assistant Professor Computer Science & Engg. Department(MANIT, Bhopal). Total 14 Years Teaching Experience (PG & UG). Guided 16 M.Tech Thesis

Mr. Bhola Nath Roy M.Tech & B.E. in Computer Science & Engg. and working as Assistant Professor Computer Science & Engg. Department(MANIT, Bhopal).