SHODH SANGAM -- A RKDF University Journal of Science and Engineering

Fuzzy -Time Correlation Method Using log file analysis and reconstruction to Understand End-User classification

Ritul Saraf , Vishal Shrivastava*

Computer Science and Engineering, Vedica Institute of Technology, Bhopal

***Abstract* - Computer crime is a crime involving computers and networks. Investigating crime in a networked environment is a tedious task. Event registration and event logs include the latest IT Critical investigation, when the last user interacts with the web environment and stores in different logs such as the client side firewall logs, gateway network logs, and server- side logs, play an important role in the system. But log files should not be stressed enough as a source of information in the system and network management. Whereas separate log files must be linked for the purpose of actively testing and collecting information. The task of parsing event log files has become inferior to continuing with the increasing size and complexity of today's logbook event. Now there is a light that one day automatically analyzes these log files. This research contains an innovative method that has been used to build a series of evidence on the basis of short and temporary series and relational algebra, and to process real generation data from logs and create Solar rules based on a number of evidence and Pre-processing classifies the actual generated data and user from the log based on the Markov model. A growing variety of data processing processes involves the audit of large log files and therefore requires processing tools and technical solutions. In the event of emergency response, human analysts have to process large amount of log data in order to detect suspicious activity and add additional evidence. In many cases, after identifying some additional facts, the reaction of this incident is stopped. Detecting online phenomena is very difficult for many reasons. Many application-specific log formats also require deep domain- specific knowledge to correctly configure an existing rules- based event-based event engine. Secondly, provide the exact model of abuse or effective identification algorithms are required.**

Keywords— Computer crime, event logs, process, event based..

I. INTRODUCTION

Event logging and event logs play an [1] important role in modern IT systems. Today, many applications, operating systems, network devices, and other system components are able to log their events to a local or remote log server. For this reason, event logs are an excellent source for determining the health status of the system, and a number of tools have been developed over the past 10-15 years for monitoring event logs in real-time. However, majority of these tools can accomplish simple tasks only, Event correlation is one of the most prominent real-time event processing techniques today. It has received a lot of attention in the context of network fault management over the past decade, and is becoming increasingly important in other domains as well, including event log monitoring. A number of approaches have been proposed for event correlation, and a number of event correlation products are available. Unfortunately, existing products are mostly expensive, platform-dependent, and heavyweight solutions that have complicated design, being therefore difficult to deploy and maintain, and requiring extensive user training. For these reasons, they are often unsuitable for employment in smaller IT systems and on network nodes with limited computing resources. So far, the rule-based approach has been frequently used for monitoring event logs – event processing tasks are specified by the human analyst as a set of rules, where each rule has the form IF condition THEN action. For example, the analyst could define a number of message patterns in the regular expression language, and configure the monitoring tool to send an SMS notification when a message that matches one of the patterns is appended to the event log. Despite its popularity, the rule- based approach has nevertheless some weaknesses – since the analyst specifies rules by hand using his/her past experience, it is impossible to develop rules for the cases that are not yet known to the analyst; also, finding an analyst with a solid amount of knowledge about the system is usually a difficult task. In order to overcome these weaknesses, various knowledge discovery techniques have been employed for event logs, with data mining methods being a common choice.

Log files are excellent sources for determining the health [2] 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.

Reconstruction of events inside a computer requires understanding of computer functionality. Many techniques emerged for reconstructing events in specific operating systems. This dissertation classifies these techniques according to the primary object of analysis. Two major classes are identified: log file analysis and file system analysis.

Digital forensics has been defined as the use of Scientifically [10] derived and proven methods towards the preservation, collection, validation, identification, analysis, interpretation and presentation of digital evidence derived from cyber sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal or helping to anticipate the unauthorized actions shown to be disruptive to planned operations One important Element of Digital forensics is the credibility of the digital evidence.

In digital forensic, log files are like the black box on an airplane 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.

PROCESS OF DIGITAL FORENSIC

To perform the digital forensic there [10] are four major steps have to use.

1 Collection

This stage consists in collecting digital source that may be relevant to the investigation process. Si digital information is stored in computers, web servers, collection of digital information means either collection of the equipment containing the information, or recording the information on some medium.

2 Examination

Examination stage consists in an in-depth systematic search of evidence" relating to the incident being investigated. The outputs of examination are data objects found in the collected information. They may include log files, data files containing specific phrases, times-stamps, and so on.

Figure 1.1 Process of Digital forensic

3 Analysis

The aim of analysis is to "draw conclusions based on evidence found".

4 Reporting

This entails writing a report outlining the examination process and pertinent data recovered from the overall investigation.

PROBLEM STATEMENT

Digital forensics [9, 10] as a discipline faces several problems, among the more acute and limiting are the following:

- Digital investigations are becoming more time consuming and complex as the volumes of data requiring analysis continue to grow.
- Digital investigators are finding it increasingly difficult to use current tools to locate vital evidence within the massive volumes of data. As a result, it is difficult for digital investigators to conduct the forensic analysis process in an effective and efficient manner, despite using state-of-the-art computer forensic tools.
- Log files are often large in size and multi-dimensional, which makes the digital investigation and search for supporting evidence more complex.

II.

OBJECTIVE OF THE DISSERTATION

Contribution of dissertation wok is to study of the field of cyber forensic, log files, role of log file in cyber forensic, evidence gathering through log file, various log files management issue and also proposes a prototype system 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.

The main approach is to correlate firewall log and web server log file for understand the end user behavior. The proposed algorithm perfume offline log files analysis by using rule based correlation and classify the suspicious user based on temporal data mining and fuzzy rule.

III LITERATURE SURVEY

1 New Approaches for Intrusion Detection Based On Logs Correlation:

Network administrators are able to correlate log file entries manually [11] Large volume and low quality of log files justify the need for further log processing. The manual log processing is lack of flexibility. It is time consuming, and one doesn't get the general view of the log files in the network. Without this general view it is hard to correlate information between the network components. Events seemingly unessential by themselves can in reality be a piece of a larger threat. In this regard, different log co relation methods are proposed to improve alert quality and to give a comprehensive view of system security.

2 Constructing Genome Phylogenetic Tree of Large dsdna Viruses Using Log Correlation Distance:

The taxonomy of the large ds DNA viruses [12] has been provided in the VIIIth report of ICTV. The phylogenetic tree of large ds DNA viruses has been constructed using CV Tree method (Gao and Qi, BMC Evol. Biol.7(2007)41). In this paper, we use the log-correlation distance method analyze the complete genome of the 124 large ds DNA viruses and construct phylogenetic trees based on compositional vectors of DNA sequences or protein sequences. The phylogenetic trees show the large dsDNA virus genomes are separated into nine families.

3 Confidentiality of event data in policy-based monitoring:

Monitoring systems observe [13] important information that could be a valuable resource to malicious users: attackers can use the knowledge of topology information, application logs, or configuration data to target attacks and make them hard to detect. The increasing need for correlating information across distributed systems to better detect potential attacks and to meet regulatory requirements can potentially exacerbate the problem if the monitoring is centralized. A single zero-day vulnerability would permit an attacker to access all information. This paper introduces a novel algorithm for performing policy-based security monitoring. We use policies to distribute information across several hosts, so that any host

compromise has limited impact on the confidentiality of the data about the overall system. Experiments show that our solution spreads information uniformly across distributed monitoring hosts and forces attackers to perform multiple actions to acquire important data.

4 A Log Correlation Model To Support The Evidence Search Process In A Forensic Investigation:

Computer forensics [14] searches for evidence to reassemble the actions that led the system from a secure state to the moment an intrusion was detected. The main source of data for a forensic investigation is the information provided by log files. Log files are generated by applications to keep a register of the actions occurred on the system. However, the massive amount of recorded events complicates the forensic investigation.

5 LEC: Log Event Correlation Architecture Based on Continuous Query:

The rapidly evolving society, every [15] corporation is trying to improve its competitiveness by refactoring and improving some if not all of its industrial software infrastructure. This goes from mainframe applications that actually handle the company's profit generating material, to the internal desktop applications used to manage these application servers. These applications often have extended activity logging features that notify the administrators of every event encounter at runtime. Unfortunately, the standalone nature of the event logging sources renders the correlation of log event infrastructure prone to continuous queries. This paper described an approach that adapts and employs continues queries for distributed log event correlation with the aim to solve problems that face the present log event management systems. It will present LEC architecture that analyze a set of distributed log events that follow a set of correlation rules; then the main output is a stream of correlated log events.

6 Log Master: Mining Event Correlations in Logs of Large-Scale Cluster Systems:

This paper [16] presented a set of innovative algorithms and a system, named Log Master, for mining correlations of events that have multiple attributions, i.e., node ID, application ID, event type, and event severity, in logs of large-scale cloud and HPC systems. Different from traditional transactional data, e.g., supermarket purchases, system logs have their unique characteristics, and hence the authors proposed several innovative approaches to mining their correlations. The authors parsed logs into an n-ary sequence where each event is identified by an informative nine-tuple. The authors proposed a set of enhanced Apriori-like algorithms for improving sequence mining efficiency, the authors proposed an innovative abstraction event correlation graphs (ECGs) to represent event correlations, and present a ECGs-based algorithm for fast predicting events. The experimental results on three logs of production cloud and HPC systems, varying from 433490 entries to 4747963 entries, show that the author's method can predict failures with a high precision and an acceptable recall rates.

IV PROPOSED ARCHITECTURE & METHODOLOGY

ALGORITHM FOR PROPOSED METHODOLOGY

Assumption

F_L = Firewall log
W_L – Web log
D_P = Distination port
R_{URI} = Restricted URI

Step 1:- Creation of sets

f_{CIP} = {CIP|CIP ∈ IP address of client in F_L between time window t_1
W_{CIP} = {CIP|CIP ∈ IP address of client in W_L between time window t_1
f_{DP} = {DP|DP ∈ distination port server between time window t_1 and

Step 2:- Creation of many to many relations

$R_{D_P} = (f_{CIP}, f_{DP})$ ie $f: f_{CIP} \rightarrow f_{DP}$
$R_{R_{URI}} = (W_{CIP}, R_{URI})$ ie $f: W_{CIP} \rightarrow R_{URI}$

Step 3:-

For(every item in f_{CIP})
{
Switch($f: f_{CIP}$)
#
$f: f_{CIP}$ return number of image by the domain (f_{CIP})of relati
over rangef_{DP}
{
Case 1 :(**$150 < f: f_{CIP} \leq 450$**)
If ($f_{CIP} \in W_{CIP}$)
{

Then if ($f: F_{CIP} \rightarrow R_{URI}$)
Go to case 2
Else
Return
F_{CIP} is less suspecious client
Go to case 2
}
Else
Return
F_{CIP} is Netwcrk suspecious client
Exit ( );
Case2: (**$350 < f: f_{CIP} \leq 550$**)
If ($f_{CIP} \in W_{CIP}$)
{
Then if ($f: F_{CIP} \rightarrow R_{URI}$)
Go to case 3
Else
Return
F_{CIP} is medium suspecious client

F_{CIP} is Netwcrk suspecious client

Exit ( );

Case3: (f : f_{CIP} >550)
If ($f_{CIP} \in W_{CIP}$)
{
Then if ($f : F_{CIP} \rightarrow K_{URi}$)
Go to case 4
Else
Return
F_{CIP} is highly suspecious client
Go to case 4
}
Else
Return
F_{CIP} is Network suspecious client
Exit ( );
Case4: (f : f_{CIP} > 200)
If ($f_{CIP} \in W_{CIP}$)
{
Then if ($f : F_{CIP} \rightarrow K_{URi}$)
Return
F_{CIP} is Attacker
Exit ( );
Else
Return
F_{CIP} is Less suspecious client
Exit ( );
}
Else
Return
F_{CIP} is Network suspecious client
Exit ( );
Case5: (f : f_{CIP} < 200)
Return F_{CIP} is normal client
}#end of switch
}# end of for

PROPOSED ARCHITECTURE

In this work there are two log files has used. One is web log and another is firewall log. But it was difficult to analyze them because of their different structure of log. Now it is necessary to convert it into the same structure. So there is a concept of temporized logging system. In this approach the log file data will convert into the database having the simple format.

There are three basic fundamentals component of the proposed work.

Temporized Logging System: This is used for the collecting data from log files with respect to time. As we know there are various log file format. This function makes the detail of log in one format in order to analysis.

Fuzzy Approach: This step is used for the applying the fuzzy association rules on the collected data. These can one or more than one rule.

Classification: Classification is a final step in order to get the suspicious users. This classification will perform on the basis of rules.

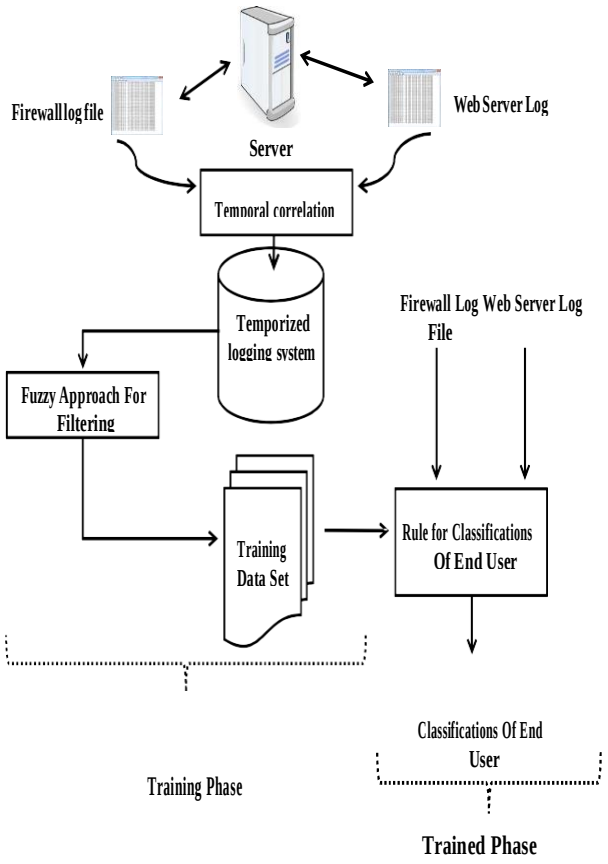

Figure: Proposed Architecture

ASSOCIATION CLASSIFICATION RULES USED

Fuzzy association classification rule (fL ,WL, fIp, fdp, nIp, HsIp, MsIp , LsIp ,swIp, atIp) having following touple
fL=firewall log file that's capture tcp/ip behaviour of network transaction
wL= Web log file that's capture http bahaviour of network transaction
fIp= finite set of client IP address of fL
fdp =finite set of Set of destination port of firewall log
Nip=finite set of normal ip
HSip=finite set of highily suspicious Ip
MSip=finite set of middium suspicious Ip
LSip=finite set of Less suspicious Ip
Swip=finite set of suspicious ip for web
ATip =finite set of attacker ip address

$\rightarrow A_{Tip}$

$fAR1: IP^{[0,10]}_{count(distinct\ fdp)>200} + IP^{[05,15]}_{Http\ transaction\ not\ having\ restricted\ zone\ http\ transaction} + IP_{having\ restricted\ zone\ http\ transaction}$

$\rightarrow Sw_{Ip}$

$fAR3: IP^{[0,10]}_{count(distinct\ fdp)>550} + IP^{[05,15]}_{Http\ transaction\ not\ having\ restricted\ zone\ http\ transaction}$

$\rightarrow HS_{Ip}$

$fAR4: IP^{[0,10]}_{350 \le count(distinct\ fdp)<550} + IP^{[05,15]}_{Http\ transaction\ Not\ having\ restricted\ zone\ http\ transaction}$

$\rightarrow MS_{Ip}$

$$fAR5: IP^{[0.10]}_{150 \le count(distinct\ fdp) < 450} + IP^{[05.15]}_{Http\ transcation} + Not\ IP_{having\ restricted\ zone\ http\ transcation} \rightarrow LS_{Ip}$$

$$fAR6: IP_{count(distinct\ fdp) < 200} \longrightarrow N_{Ip}$$

Where *fAR1* represent the scenario to cache the attacker i.e. if any IP address performs the port scanning and having number of port scan is greater than 200 along with that it have HTTP transaction between 05 to 15 unit time offer port scan and it was try to enter in restricted zone of web server admin then it is an attacker.
Where *fAR2* show the scenario for suspicious IP, IP that perform total number of port scan is greater than 200 but having no HTTP transaction is suspicious for network not for web.

Where *fAR3, fAR4, fAR5* represent the degree of suspicious user for web i.e. *fAR3* represent highly suspicious user

Where *fAR4* represent medium suspicious user and *fAR5* represent low suspicious user with various number of ports scan. In *fAR3, fAR4, fAR5* there is overlap of boundary for number of port scan i.e. fuzzy approach. In last *fAR6* represent the scenario for normal users.

All the rules namely *fAR1* to *fAR6* is time based.

V .SIMULATION ENVIRONMENT & RESULT ANALYSIS

The implementation of the proposed work has done in MATLAB 10.0 with the help of MySQL database. This work we uses Intel core i3 CPU with 2.53 GHz, having 4 GB of RAM and 500 GB HDD. We worked on 32 bit OS (Windows 7).

1. SCREEN SHOTS

In this section we attached the various screen shots in order to show the outputs of the proposed work. As we discussed about the environment of implementation these outputs are generated by the MATLAB.
A client server architecture having many clients and one server is taken as a scenario for verification of proposed work whole verification is done over MATLAB 7.10.

log file i.e. firewall log file is responsible for TCP behavior and the server side log file is responsible for HTTP behavior in log file.

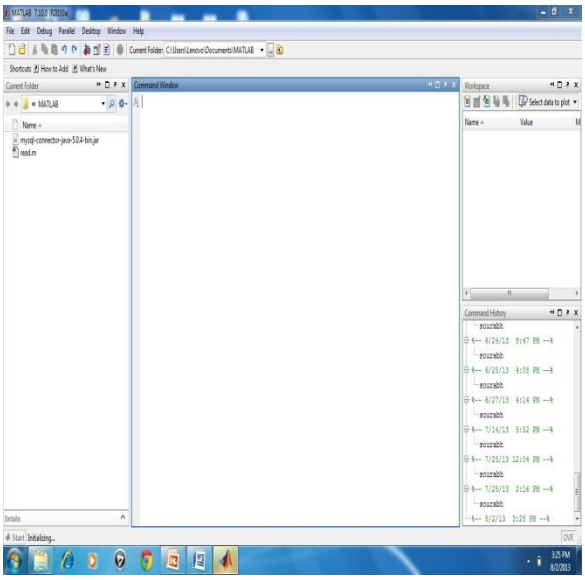

Figure: Initial Stage

Where as client server architecture responsible for generating log file at both client and server side. Client side

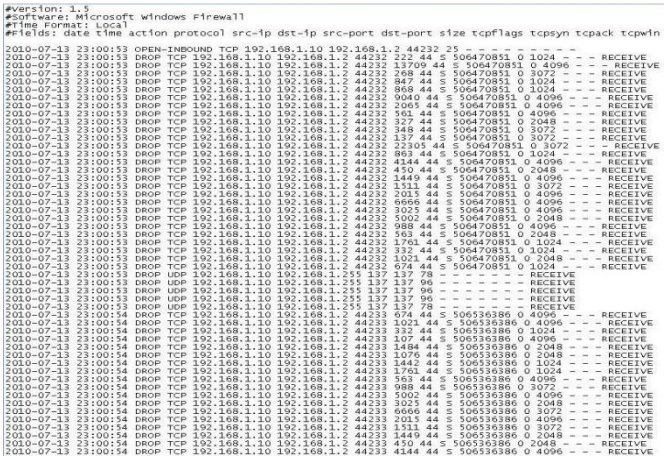

Figure: Firewall log file

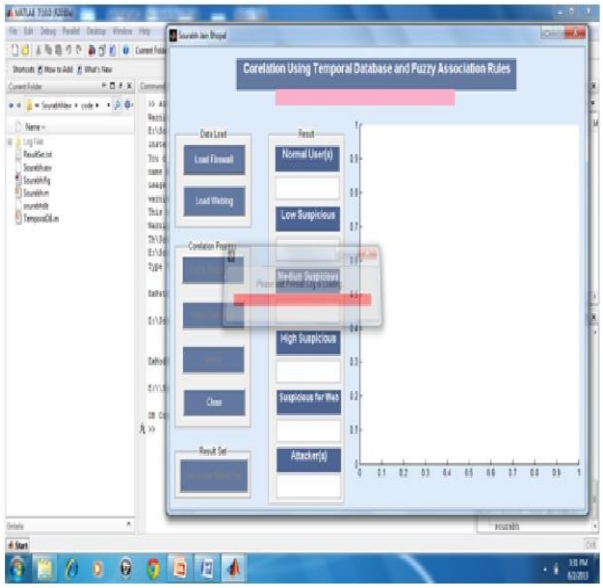

Figure: Firewall log loading

Above figure shows the process of firewall log file loading from the database this log is a sample training data set. It contains information about all the incoming and outgoing activities in the network.

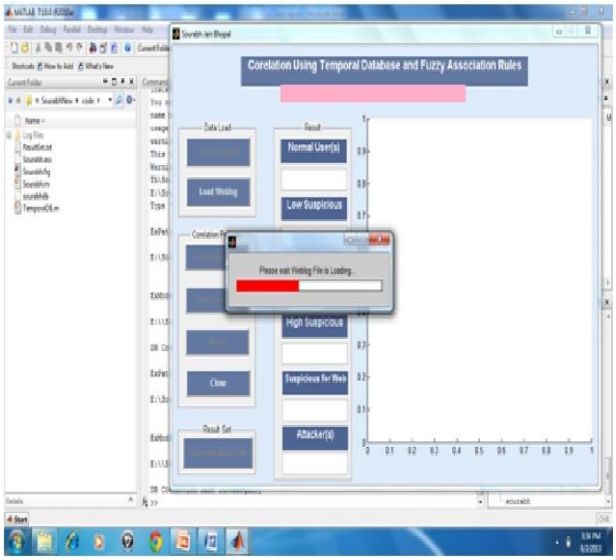

Figure: web log loading

Above figure shows the process of web log file loading from the database this web log is a sample training data set. This sample data set contains information about Users activities from server, Protocol status, file transfer etc.

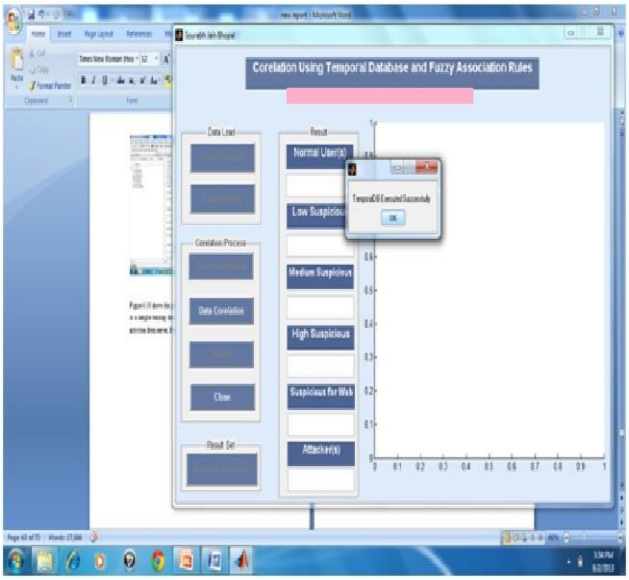

Figure: Fuzzy Association Rule Executed

Figure shows that after uploading web log and firewall files the fuzzy association rule is executed successfully the fuzzy association rules are used for the classification of data. These rules can applicable where there is some probability. This method is efficient to get the results from the boundary cut problems.

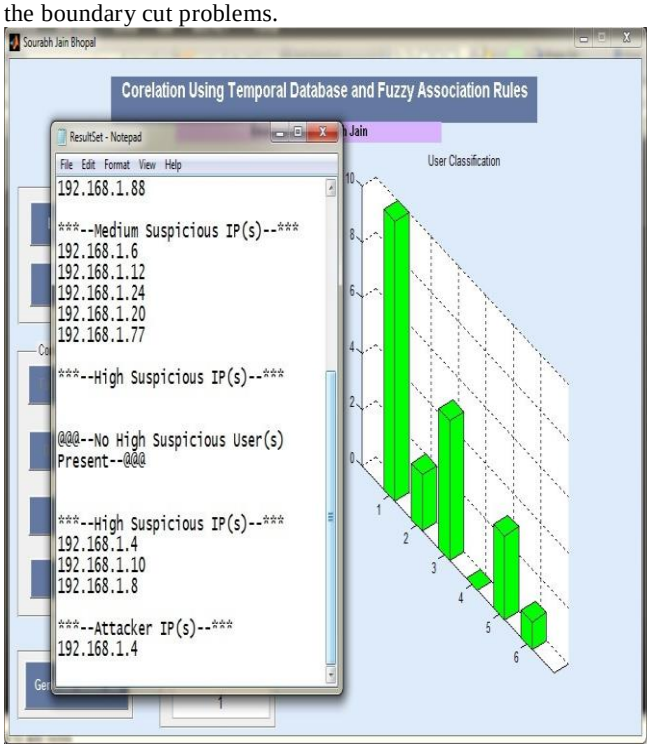

Figure: Result data Set

Above figure shows result dataset this result data set is training data set generated by
Integrating both the log files (web log & firewall log) this resultant data set contains
IP address, web entry, restricted zone etc.

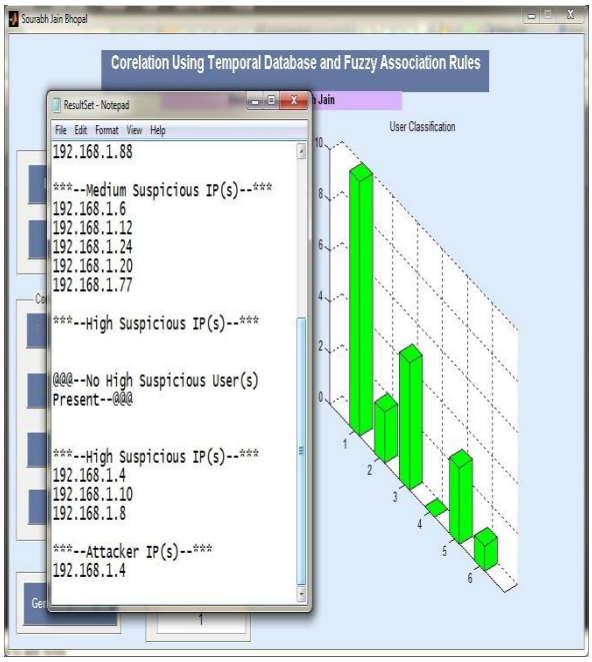

Figure: Final Result

Figure shows Proposed methodology correlate both HTTP and TCP behavior of log file by using fuzzy temporal association rule and classify the end user wither the end users are normal, suspicious, attackers.

2. COMPARISON TABLE

In this dissertation, the MATLAB simulated experiments are performed to verify the accuracy of proposed model. Log format synchronization is one of great challenge in log management issue, recently researcher focus on that problem. Proposed model in [1, 2] is log format dependent where as Proposed model is not format dependent

Along with that log rotation (size of log file) and clock synchronization is another most challenging issue in log management .Proposed model is time dependent and result set is independent form size of log file.

V. CONCLUSION AND FUTURE WORK

CONCLUSION

These days the level of computer crime has increased dramatically. We need to improve the investigative system to identify the culprit. Web server logs are usually fixed on the behavior of the machine, and not on the behavior of the end user. The log file provides troubleshooting, security, and proactive system administration, which provides significant support for suspicious users in the caching and cyber expertise process. In this dissertation, implemented system extracts the evidence from log file and correlates these generated logs on the basis of relational algebra and classifies end user .v model. We have proposed a novel log analysis method using TL based on reconstruction. The proposed method provides a solution to identify an attacker. This approach uses time-based data analysis and fuzzy communication rules. As the results show, the proposed thesis provides better results which are better than the previous work. The proposed baseline work encourages a web researcher to navigate the end-user behavior and ensure an effective security policy.

FUTURE SCOPE

Although the initial results are encouraging, there is still a lot of work to do for improving the evidence gathering efficiency. A major issue's for future work are related to log consistency, log integrity and log rotation to do more extensive test with large volume of log data. Future work also includes clock synchronization problem arises in evidence correlation.

REFERENCES

[1]. Risto Vaarandi "Tools and Techniques for Event Log Analysis", Faculty of Information Technology, Department of Computer Engineering, Chair of System Programming, Tallinn University of technology,2005

[2]. 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.

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

[4] Spafford, E.H "Defining Digital Event Reconstruction" Journal of sciences, 49(6). Paper ID 004
, "Application Of Formal Methods se Analysis of Digital Incidents", ournal of Digital Evidence, 3(1)

[5]. Bennie Kar Leung Fei, " Data Visualisation In Digital Forensics" University of Pretoria etd-Fei,BKL,2007

[6]. S.O. Azarkasb, S.S. Ghidary, "New Approaches for Intrusion Detection Based On Logs Correlation" IEEE 2009, pp 234.

[7]. L.Q.Zhou and J.M.Bai, "Constructing Genome Phylogenetic Tree of Large dsdna Viruses Using Log-Correlation Distance", IEEE 2010, pp 2182-2185.