IOSR Journal of Engineering (IOSRJEN)
e- ISSN: 2250-3021, **p**-ISSN: 2278-8719
Vol. 3, Issue 7 (July. 2013), ||V5 || PP 11-15

Detection of Abnormal Attributes Based on Corelation Markov Detection Method For Forensic Web

Sujeet Singh, Jitendra Jatav, Gourav Shrivastava.
Department of Computer Science and Engineering RKDFIST, BHOPAL

Abstract: - Crime investigation in the web environment is a tedious job. Event logging and event logs play an important role in modern IT systems criminal investigation which is generated when end user with each other in web environment and stored in various logs like firewall log file at side ,network log file at gateway and web log file at server side. But log file is not to be over emphasized as a source of information in systems and network management. Whereas conduct efficient investigation and gathering of use full information need to correlate different log file. Task of analyzing event log files with the ever-increasing size and complexity of today's event logs has become cumbrous to carry out manually. Nowadays latest spotlighted is automatic analysis of these logs files. In this paper a novel methodologies based on relational algebra to build the chain of evidence and used to preprocess the real generated data from logs and classify the user based on markov model.

Keywords: - *cyber forensic; log file correlation; markov chain ,cyber crime;*

I. INTRODUCTION

As digital crimes continue to rise, the need for digital forensics also increases. Digital forensics is utilized to conduct investigations into digital crimes or incidents. The aim of such investigations is to expose and present the truth, which often leads to prosecution and conviction. Dramatic increases in the numbers of digital crimes committed have led to the development of a whole slew of computer forensic tools. These tools ensure that digital evidence is acquired and preserved properly and that accuracy of results regarding the processing of digital evidence is maintained [1]. Researchers that might benefit from research in this area are those involved in network administration. And small to medium sized companies may not have resources to outsource this part of their system due to economic considerations. Such companies can benefit from calculating saved resources of implementing a system [2]. Benefits of such a rationalization and increased security are points that are easily supported by the management. The relation between expected benefits can be represented like this: centralized analysis to log files → easily survey able → Better detection of correlating events → rationalization in time consumption of log file examination/cost effectiveness → quicker response time → increased security[1,3].

The research described in this paper focuses on 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 end user behavior on the basis of Markov model.

II. PROPOSED FRAMEWORK

Figure 1 shows proposed architecture of prototype system. Presented architecture is five layer's architecture where duty of each and every layer is exclusive but depends upon previous layer output. In this architecture first preprocess the real generated data (record of event perfume by each and every client) from logs then passing its via whole layered processor at last classify user on the basis of Markov Model. The Proposed framework can be defined by following stages.

A. Centralization of Log files:

In this step, log files maintained by the web server and firewall are extracted to be to be centralized on the centralization server. The files are converted in such a format that analysis work can be done easily. The centralization server may have any compatible database as a backend, which can store numerous entries as they are.

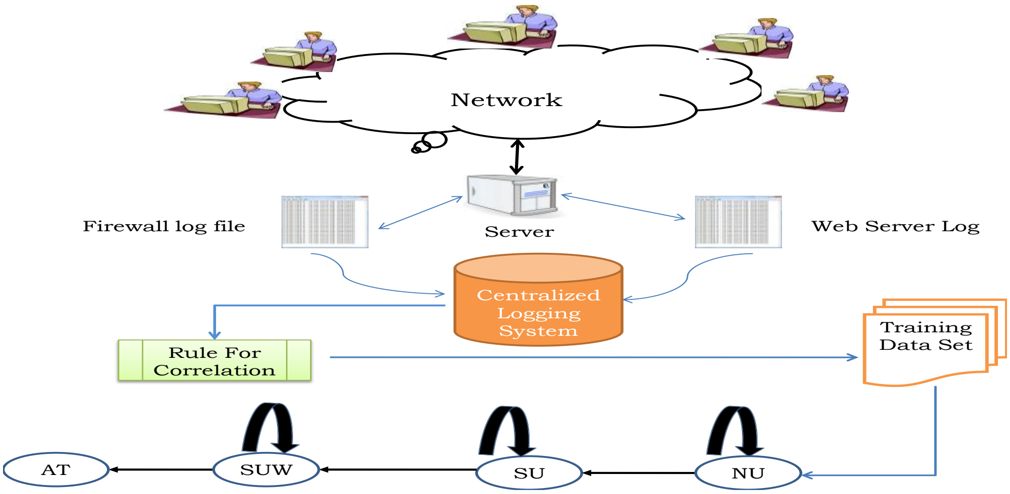

Figure 1: Proposed frame work for client classification using Markov Model

B. Correlation:

Correlation is the process of analysing and determining a set of related events, based on a set of rules that are used to interpret the data contained in the events [4]. There are several types' correlations; some of them are as follows,

- Sequential correlation: The ability to sort the events in a log by various fields contained in the events (for example, time stamp)
- Associative correlation: The ability to filter or group the events displayed in a log by the values in various fields contained in the events (for example, grouping by thread ID). The correlation types can be used together to provide a complete picture. For example, when grouping a set of events together you typically also order the events in the group.

C. Markov Model:

The Markov Model is a powerful statistical tool in order to modelling generative sequences that can be characterized by an underlying process generating an evident sequence. In other words A Probabilistic process over a finite set is known as Markov model. These final set called its states for a instance {S1,S2,….Sk}. [5]

III. WORKING MODEL

In Proposed Model for a prototype system has been created and implemented, which takes firewall log and the web log of same time as input and generates a resultant training data set, which is used as training data set for decision tree construction where decision tree create classification rule regarding client Behavior whether client is normal user, suspicious user, suspicious web user or an attacker.

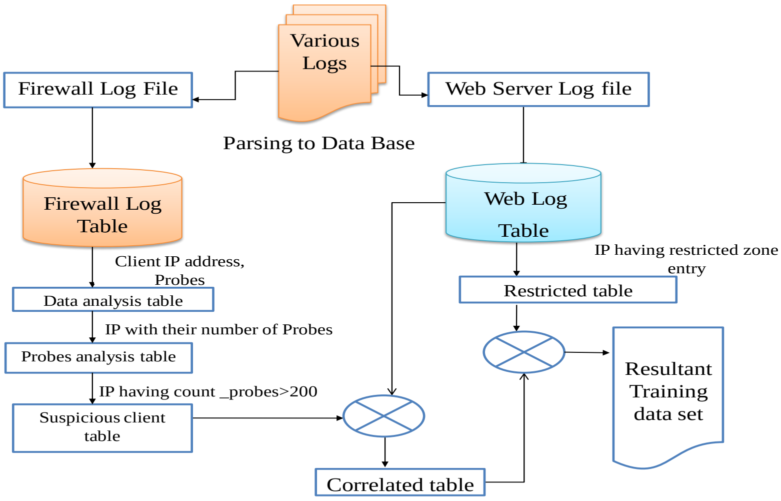

Figure 2: Working Model of Proposed Work

IV. IMPLEMENTATION DETAILS

Proposed frame work has been implemented by using a real time scenario of client server architecture having 20 clients and 1 server. Which capture firewall log and the web log file at client side and server side respectively of same time and use it as input and generates a Markov model that analysis client behavior whether client is normal user, suspicious user, suspicious web user or an attacker?

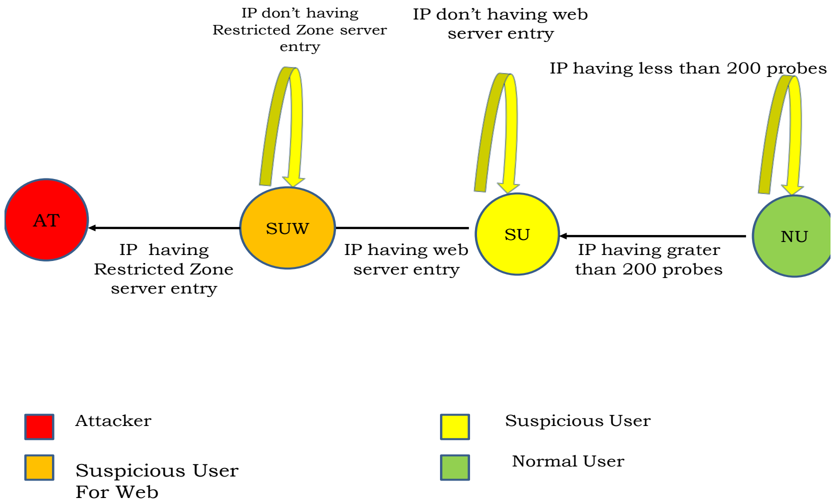
IP don't having Restricted Zone server entry
IP don't having web server entry
IP having less than 200 probes
AT
SUW
SU
NU
IP having Restricted Zone server entry
IP having web server entry
IP having grater than 200 probes
🟥 Attacker
🟨 Suspicious User
🟧 Suspicious User For Web
🟩 Normal User

Figure 3: Resultant Markov Model Extracting of assoisation Classification rules based on markov model:

Markov Model shows in figure figure 5 behave like Moore machine can be classified the end user by each state.

- IF Client_IP having probes >=200 U Client_IP having weblog file entry = "yes" U Client_IP having restricted_zone entry = "yes" then Client_IP = "Attacker"
- IF Client_IP having probes >=200 U Client_IP having weblog file entry = "yes" U Client_IP having restricted_zone entry = "No" then Client_IP = " Suspicious user for web"
- IF Client_IP having probes >=200 U Client_IP having weblog file entry = "No" then Client_IP = " Suspicious user for web"
- IF Client_IP having probes <=200 U Client_IP having weblog file entry = "yes" U Client_IP having restricted_zone entry = "No" then Client_IP = " Normal User"

V. RESULT ANALYSIS

In this paper, the MATLAB (7.1.4) 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. For the performance evaluation of proposed model we used two logfiles one is firewall log file and another one is web log file. Both log files we get by UCI machine learning site for research purpose. UCI site is famous and well know website provide data for educational research purpose. The total number of log enter is 67168 from firewall and 326 from web log. In this evaluation we detected the number of abnormal user and correlation suspicious time and error rate of detection in exiting method and our proposed method. For the better performance of result we create four different dataset in combination of two log files. The evaluation result table given below

Metric		Error rate (%)	Precision (%)	Recall (%)
Data-Set 1	DCD	3.14	88.56	85. 34
	CMD	2.14	97.32	95.21
Data-Set 2	DCD	9.90	84.32	83.23
	CMD	5.23	92.14	91.21
Data-Set 3	DCD	1.34	86.14	85.11
	CMD	5.12	93.21	91.13
Data-Set 4	DCD	2.22	88.21	87.66
	CMD	1.13	94.52	93.67

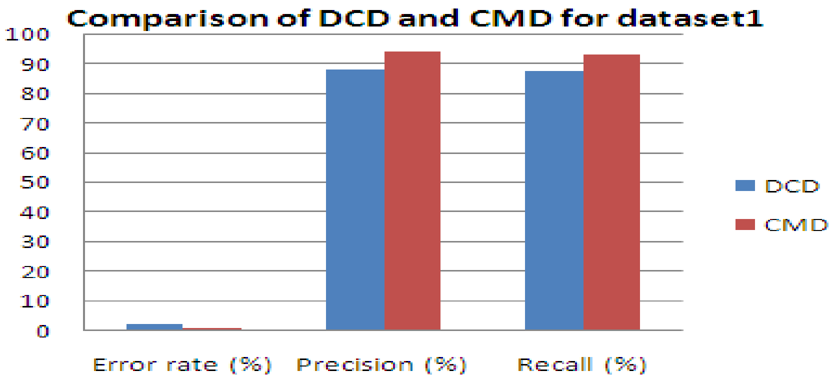

Figure 4.1 shows that the performance of dataset 1 in this dataset the total number of log entries are67494.

The processing of data measure the precision and recall as well as error rate of confusion matrix during find the similar and dissimilar pattern.

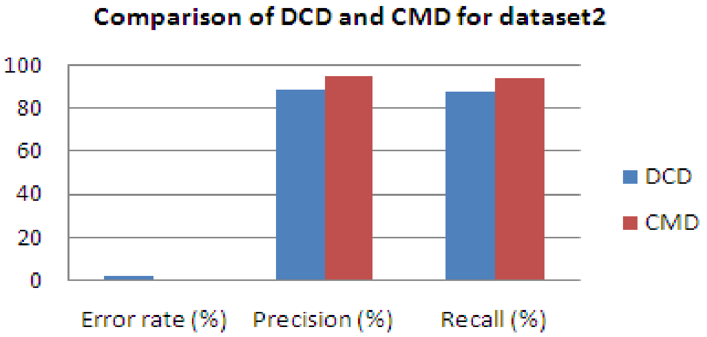

Figure 4.2 shows that the performance of dataset 2 in this dataset the total number of log entries are67368.

The processing of data measure the precision and recall as well as error rate of confusion matrix during find the similar and dissimilar pattern.

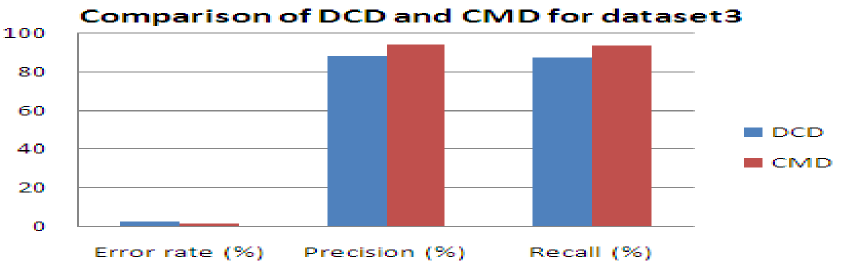

Figure 4.3 shows that the performance of dataset 3 in this dataset the total number of log entries are67326.

The processing of data measure the precision and recall as well as error rate of confusion matrix during find the similar and dissimilar pattern.

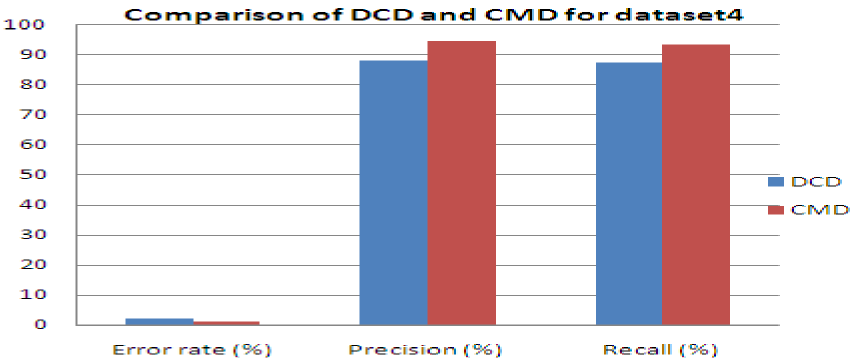

Figure 4.4 shows that the performance of dataset 4 in this dataset the total number of log entries are60326.

The processing of data measure the precision and recall as well as error rate of confusion matrix during find the similar and dissimilar pattern.

VI. CONCLUSIONS

Proposed model extracts the evidence from log file and correlate these generated logs on the basis of relational algebra and classifies end user on the bases of Markov model. Proposed frame work encourages the web investigator to navigate the end user behavior and assist to enforce the effective security policy. The future work will cover the issues related to log consistency, log integrity and other log management issue.

VII. ACKNOWLEDGMENTS

The research presented in this paper would not have been possible without our college, at RKDF IST 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 it is. We also wish to thank the anonymous reviewers for their valuable suggestions.

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]. Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey,ACM SIGKDD Explorations Newsletter, June 2000, Volume 2 Issue 1.

[4]. L. Baum et. al. A maximization technique occuring in the statistical analysis of probablistic functions of markov chains. Annals of Mathematical Statistics, 41:164–171, 1970.

[5]. Katherine A.Heller,YEE Whye The And Dilan , "Infinite Hierarchical Hidden Markov Models" in Proceedding Of the12th International conference on AISTATS,2009

[6]. Stefan Hommes, Radu State, Thomas Engel, "A Distance-Based Method to Detect Anomalous Attributes in Log Files" in IEEE ,2012

[7]. P.W.D.C. Jayathilake, "A Novel Mind Map Based Approach for Log Data Extraction " in 6TH international conference on industrial and information system,IEEE,,2011

[8]. Thomas Reidemeister, Miao Jiang and Paul A.S. Ward, "Mining Unstructured Log Files for Recurrent Fault Diagnosis" in IEEE,2011