Identifying Network Anomalies Using Clustering Technique in Weblog Data

B. Kiran Kumar
Associate Professor
Department of M.C.A.
Kakatiya Institute of Technology & Science,
Warangal, A.P., INDIA

A. Bhaskar
Associate Professor
Department of M.C.A.
Kakatiya Institute of Technology & Science,
Warangal, A.P., INDIA

ABSTRACT

In this paper we present an approach for identifying network anomalies by visualizing network flow data which is stored in weblogs. Various clustering techniques can be used to identify different anomalies in the network. Here, we present a new approach based on simple K-Means for analyzing network flow data using different attributes like IP address, Protocol, Port number etc. to detect anomalies. By using visualization, we can identify which sites are more frequently accessed by the users. In our approach we provide overview about given dataset by studying network key parameters. In this process we used preprocessing techniques to eliminate unwanted attributes from weblog data.

Keywords

Network flow data, Anomalies, Clustering, Visualization.

1. INTRODUCTION

As the increase in demand of using Internet, networks are growing rapidly and became open to public access. This leads to increase in the number and type of intrusions dramatically. This poses serious challenges for network analysis to detect various anomalies occur at different network proximities. We can use the firewalls to prevent illegal accessing of web pages. But, network intrusions take an advantage of vulnerabilities in some systems in the computer networks. Hence, there is a need for intrusion detection system beside firewalls to track intrusions in our network. Intrusion detection systems are mainly of two types: 1) Anomaly Detection System (ADS): It shows the normal behavior of network or user or application and identifies deviations to these profiles which may be potential security breaches. 2) Misuse Intrusion Detection System (MIDS): It uses attack signatures to compare with packet payloads for identifying intrusions. Using this method, it is not possible identify new attacks.

Now-a-days, speed, complexity and the size of the network is growing rapidly, and the networks became open to public access, there is a tremendous increase in number and type of intrusions so that making it impossible for human analysis. To make analysis simple we can use different data mining techniques for network intrusion detection and analysis of network flow data. There are various approaches that use data mining techniques such as fuzzy logic [1], neural networks [2] and agent based [3] data mining approaches are widely used in intrusion detection systems. Clustering, association rule and sequential association rule mining are well known data mining techniques used in intrusion detection systems. Clustering technique partitions the data items in to finite number of groups based on their similarity. Partitioning Methods, Hierarchical Methods, Density-Based Methods, Grid-Based Methods, Model-Based Clustering Methods are different clustering techniques available in data mining. In this paper we are using Simple K-Means clustering technique to partition the weblog data to detect the network anomalies. In applying visualization to internet security, researchers exploit the human ability to process visual information quickly enables the complex task of network security monitoring and intrusion detection to be performed accurately and efficiently as discussed in [4][5]. In this paper we collected network flow data from the web server of our organization, after which we carryout preprocessing and filtering of the data. We partitioned the preprocessed data to form clusters based on various attributes of data set. We used JFree Chart plug-in software for JAVA to visualize the clusters. By cluster analysis we can identify different anomalies in the data set.

The rest of the paper is organized as follows. In section 2, we present the related work and discuss various techniques used for visualization. We next present our approach in section 3. In section 4 we present results and analysis. Finally, conclusions are made in section 5.

2. RELATED WORK

To solve a number of problems encountered by the Intrusion detection, Visualization is a technique used in intrusion detection system. Conti et. a1 [4] provided a survey of packet and alert visualization techniques and present the challenges involved in information visualization of security related data and present techniques for network traffic visualization. D'Amico et.al [6] presented a technique for visualization that depicts patterns in massive amounts of data, and present methods for interacting with those visualizations to help analysts prepare for unforeseen events. Xiao et.al [7], present their work on visualization of network traffic that is applied for classifying the traffic and used a PLOT to visualize the data set. Itoh et.al[8] presented different techniques for visualizing the content of huge IDS log files. Goodall et.al [9], present the technique called as Time Based Network Traffic Visualization using which the complex task of searching for indications of attacks and misuse in vast amounts of network data is carried out. Tee et.al [10] share their work based on the Origin Autonomous System Change technique which is based on the premise that we can glean valuable knowledge from large data sets the same premise behind knowledge discovery. Munz et.al [11] presented a flow based anomaly detection scheme based on K-Means clustering algorithm where they cluster the unlabelled flow data for normal and anomalous traffic, but our approach is focused to provide intuitive visualization of network traffic based on flow analysis by applying Simple K-Means Clustering Algorithm.

Visualization Techniques:- Pen et. al [12] discusses that how intrusions can be detected by visualizing the cluster groups through a technique called IDGraphs. Conti et.al[4]presents a technique called ID RAINSTROM for visualizing the data. D’Amico et.al[6] presents a technique Vi-Assist, Abdullah et.al[5]discusses about scaling technique and Stacked Histogram. Goodall et.al[9] used Time Based Network Traffic visualization.

3. OUR APPROACH

In this paper we discuss the data collection method, preprocessing and filtering of the data. Then we used K-Means clustering algorithm to form the clusters on different network flow attributes. After this we analyzed those clusters by visualizing the flow data to detect the network anomalies. Figure.1 depicts the overall process of our approach. First we collect network flow data in the form of web log records from web server. The sample of collected raw data is shown in figure 2.

Fig 1: Process Flow

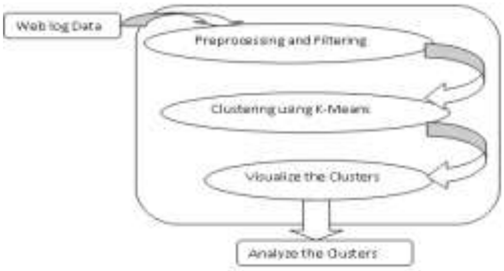
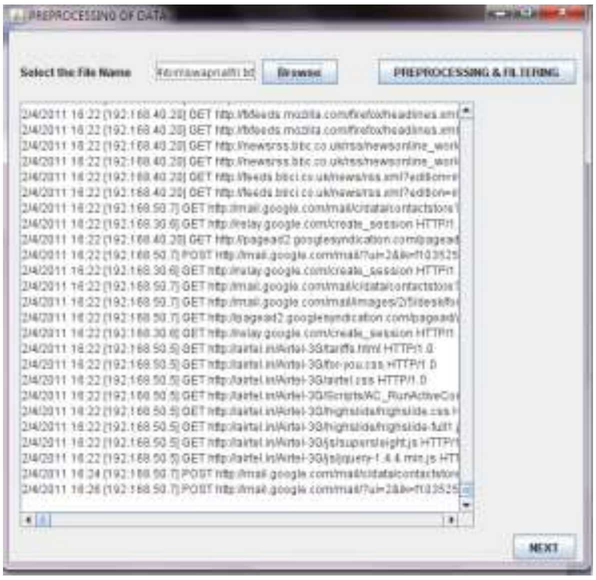


Fig 2:Raw Web log data

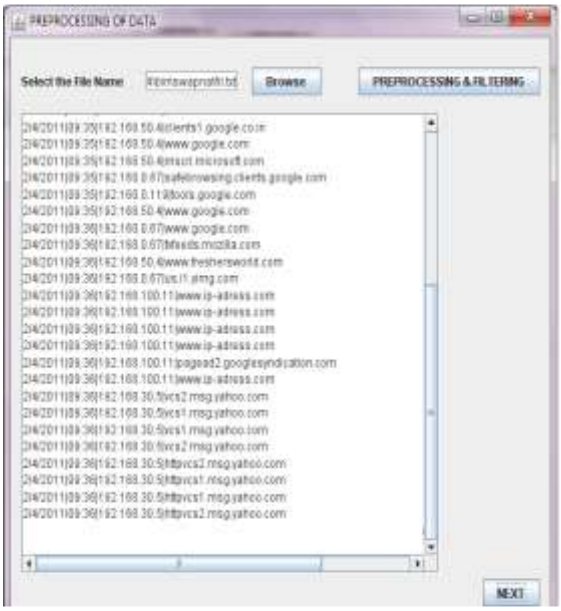


Select the File Name #bimaswaprathi.txt

```
2/4/2011 16:22 [192.168.40.20] GET http://fxfeeds.mozilla.com/firefox/headlines.xml
2/4/2011 16:22 [192.168.40.20] GET http://fxfeeds.mozilla.com/firefox/headlines.xml
2/4/2011 16:22 [192.168.40.20] GET http://newsrss.bbc.co.uk/rss/newsonline_worl
2/4/2011 16:22 [192.168.40.20] GET http://newsrss.bbc.co.uk/rss/newsonline_worl
2/4/2011 16:22 [192.168.40.20] GET http://feeds.bbci.co.uk/news/rss.xml?edition=i
2/4/2011 16:22 [192.168.40.20] GET http://feeds.bbci.co.uk/news/rss.xml?edition=i
2/4/2011 16:22 [192.168.50.7] GET http://mail.google.com/mail/c/data/contactstore
2/4/2011 16:22 [192.168.30.6] GET http://relay.google.com/create_session HTTP/1
2/4/2011 16:22 [192.168.40.20] GET http://pagead2.googlesyndication.com/pagead
2/4/2011 16:22 [192.168.50.7] POST http://mail.google.com/mail/?ui=2&ik=f103525
2/4/2011 16:22 [192.168.30.6] GET http://relay.google.com/create_session HTTP/1
2/4/2011 16:22 [192.168.50.7] GET http://mail.google.com/mail/c/data/contactstore
2/4/2011 16:22 [192.168.50.7] GET http://mail.google.com/mail/images/2/5/deskfo
2/4/2011 16:22 [192.168.50.7] GET http://pagead2.googlesyndication.com/pagead/
2/4/2011 16:22 [192.168.30.6] GET http://relay.google.com/create_session HTTP/1
2/4/2011 16:22 [192.168.50.5] GET http://airtel.in/Airtel-3G/tariffs.html HTTP/1.0
2/4/2011 16:22 [192.168.50.5] GET http://airtel.in/Airtel-3G/for-you.css HTTP/1.0
2/4/2011 16:22 [192.168.50.5] GET http://airtel.in/Airtel-3G/airtel.css HTTP/1.0
2/4/2011 16:22 [192.168.50.5] GET http://airtel.in/Airtel-3G/Scripts/AC_RunActiveCo
2/4/2011 16:22 [192.168.50.5] GET http://airtel.in/Airtel-3G/highslide/highslide.css H
2/4/2011 16:22 [192.168.50.5] GET http://airtel.in/Airtel-3G/highslide/highslide-full.
2/4/2011 16:22 [192.168.50.5] GET http://airtel.in/Airtel-3G/js/supersleight.js HTTP/1
2/4/2011 16:22 [192.168.50.5] GET http://airtel.in/Airtel-3G/js/jquery-1.4.4.min.js HTT
2/4/2011 16:24 [192.168.50.7] POST http://mail.google.com/mail/c/data/contactstore
2/4/2011 16:26 [192.168.50.7] POST http://mail.google.com/mail/?ui=2&ik=f103525
```

This data set is preprocessed and then filtered based on key parameters like IP address, Protocol, and URL. The preprocessing step is very much important for our approach since it decides the formation of clusters. In preprocessing step we trimmed the length of the URL by eliminating unnecessary control characters, and request parameters. Preprocessing is done in such a way that the clusters do not overlap and distant apart. After the preprocessing the data set is filtered by eliminating unwanted attributes. In the filtering step we eliminated request method (GET/POST). The preprocessed and filtered data set is shown in figure 3. After filtering of the data set, we applied the simple K-Means algorithm on some meaningful attributes like IP address, Protocol and URL. The resultant clusters are shown in figure 4.

Fig 3: Preprocessed and Filtered data.

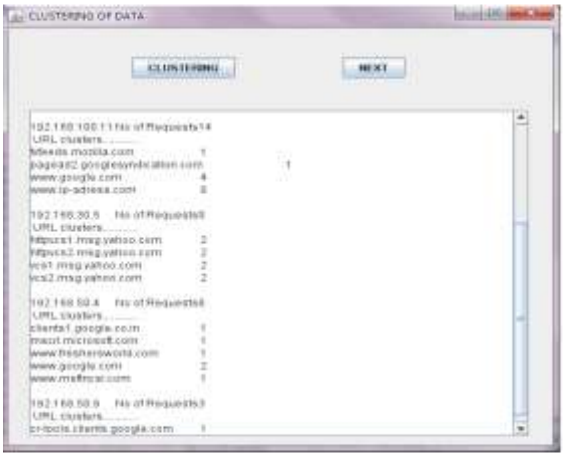


Select the File Name #bimaswaprathi.txt

```
2/4/2011|09:35|192.168.50.4|clients1.google.co.in
2/4/2011|09:35|192.168.50.4|www.google.com
2/4/2011|09:35|192.168.50.4|msoct.microsoft.com
2/4/2011|09:35|192.168.0.67|safebrowsing.clients.google.com
2/4/2011|09:35|192.168.0.119|tools.google.com
2/4/2011|09:35|192.168.50.4|www.google.com
2/4/2011|09:36|192.168.0.67|www.google.com
2/4/2011|09:36|192.168.0.67|fxfeeds.mozilla.com
2/4/2011|09:36|192.168.50.4|www.freshersworld.com
2/4/2011|09:36|192.168.0.67|us.i1.yimg.com
2/4/2011|09:36|192.168.100.11|www.ip-adress.com
2/4/2011|09:36|192.168.100.11|www.ip-adress.com
2/4/2011|09:36|192.168.100.11|www.ip-adress.com
2/4/2011|09:36|192.168.100.11|www.ip-adress.com
2/4/2011|09:36|192.168.100.11|pagead2.googlesyndication.com
2/4/2011|09:36|192.168.100.11|www.ip-adress.com
2/4/2011|09:36|192.168.30.5|vcs2.msg.yahoo.com
2/4/2011|09:36|192.168.30.5|vcs1.msg.yahoo.com
2/4/2011|09:36|192.168.30.5|vcs1.msg.yahoo.com
2/4/2011|09:36|192.168.30.5|vcs2.msg.yahoo.com
2/4/2011|09:36|192.168.30.5|httpvcs2.msg.yahoo.com
2/4/2011|09:36|192.168.30.5|httpvcs1.msg.yahoo.com
2/4/2011|09:36|192.168.30.5|httpvcs1.msg.yahoo.com
2/4/2011|09:36|192.168.30.5|httpvcs2.msg.yahoo.com
```

The formed clusters are visualized and analyzed based on IP address, Protocol and URL to identify network anomalies.

Fig 4: Clusters



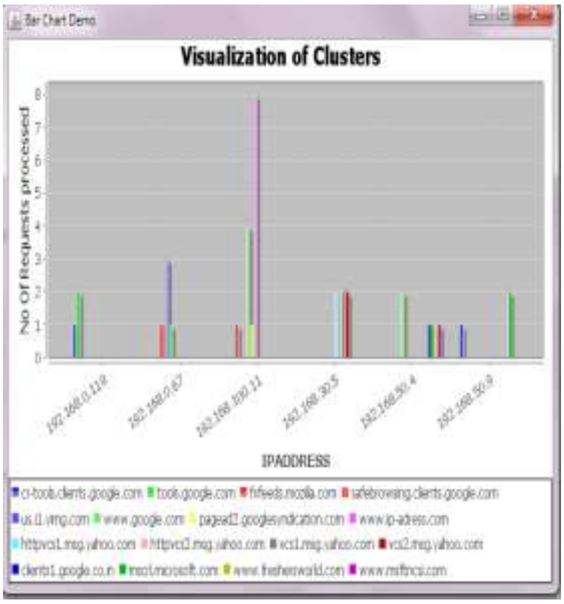
```
192.168.100.11 No of Requests14
 URL clusters..........
fxfeeds.mozilla.com              1
pagead2.googlesyndication.com              1
www.google.com                   4
www.ip-adress.com                8

192.168.30.5     No of Requests8
 URL clusters..........
httpvcs1.msg.yahoo.com           2
httpvcs2.msg.yahoo.com           2
vcs1.msg.yahoo.com               2
vcs2.msg.yahoo.com               2

192.168.50.4     No of Requests6
 URL clusters..........
clients1.google.co.in            1
msoct.microsoft.com              1
www.freshersworld.com            1
www.google.com                   2
www.msftncsi.com                 1

192.168.50.9     No of Requests3
 URL clusters..........
cr-tools.clients.google.com      1
```

Fig 5: Visualization of Clusters.



4. RESULTS AND ANALYSIS

We used General Public License (GPL) open source software JFree Chart for JAVA to visualize the formed clusters. First we applied clustering on IP address, where we obtained 6 clusters. On these clusters we again performed clustering on URL. In the analysis of protocol attribute we observed that, in case of normal flow data TCP and UDP packets are almost equal in number. But, in case of attack flow data TCP packets are more in number compared to UDP, few ICMP and IGMP packets also be observed. With this protocol attribute we can detect attacks like Denial of Service (DoS) and malware spreading. The visualization chart is shown in figure 5.

5. CONCLUSIONS

In this paper we presented an approach for analyzing and visualizing the network flow data using clustering. It is an easy, simple and fast way of analyzing the flow data. By the help of clustering we can predict the type of flow i.e. attacks or normal by performing some clustering on the particular attributes. We presented our analysis mainly based on three attributes. In this paper we have performed operation only on limited number of attributes. We can increase the number of attributes to be analyzed which shall give a much clear picture of the type of data. Also the algorithm of the preprocessing can be enhanced further.

5. REFERENCES

[1] Hai Jin, Jianhua Sun, Hao Chen, Zongfen Han. "A Fuzzy Data Mining Based Intrusion Detection Model ", *Proceedings of the 10th IEEE International Workshop on Future Trends of Distributed Computing Systems, 2004.*

[2] John Zhong Lei, Ali Ghorbani, "Network Intrusion Detection using Improved Competitive Learning Neural Network.", *Proceedings of the Second Annual Conference on Communication Networks and Services Research, 2004.*

[3] Cheung Leung Lui, Tak Chung Fu ,Ting Yee Cheung. "Agent-based Network Intrusion Detection System Using Data Mining Approaches ", *Proceedings of the Third International Conference on Information Technology and Applications, 2005*.

[4] Gregory Conti, Kulsoom Abdullah, Julian Grizzard, John Stasko, John A. Copeland, Mustaque Ahamad, Henry L. Owen, and Chris Lee. "Countering Security Information Overload through Alert and Packet Visualization", Published *by the IEEE Computer Society, March/April 2006.*

[5] Kulsoom Abdullah, Chris Lee, Gregory Conti, John A. Copeland. "Visualizing Network Data for Intrusion Detection", *Proceedings of the IEEE 2002.*

[6] D. D'Amico, John R. Goodall, Daniel R. Tesone, Jason K. Kopylee. "Visual discovery in computer network" published by the IEEE computer Society Sept/Oct, 2007.

[7] Ling Xiao,John Gerth,Pat Hanrahan. "Enhancing Visual Analysis of Network Traffic Using Knowledge Representation ",*Proceedings of the IEEE 2006.*

[8] Takayuki Itoh,Hiroki Takakura,Atsushi Sawada and Koji Koymada. "Hierarchical Visualization Of network Intrusion Detection Data",*Proceedings of the IEEE 2006.*

[9] John R. Goodall, A. Ant Ozok, Wayne G. Lutters, Penny Rheingans, Anita Komlodi. "A User-Centered Approach to Visualizing Network Traffic for Intrusion Detection ",*Proceedings of the ACM, APRIL 2,2004,USA*

[10] Soon tee, Teoh, Kwan-Liu, Ma,Soon Felix Wu,T.J. "Detecting Flaws and Intruder with Visual Data Analysis ", Published by the IEEE Computer Society, IEEE Computer Graphics and Applications 2004.

[11] Gerhard Münz, Sa Li, and Georg Carle. Traffic anomaly detection using k-means clustering. In Proc.of Leistungs-, Zuverlässigkeits- und Verlässlichkeitsbewertung von Kommunikationsnetzen und Verteilten Systemen, 4.GI/ITG-Workshop MMBnet 2007, Hamburg, Germany, September 2007.

[12] Pin Pen,Yan Gao,Zhichun Li,Yan Chen,Benjamin Watson, "IDGraph :Intrusion Detection and analysis using Stream Compositing" ,*Published by the IEEE Computer Society, March/April 2006.*