Research Article Impact Factor: 4.226 ISSN: 2319-507X
Deepti Vishnu Patange, IJPRET, 2015; Volume 3 (9): 626-633 IJPRET


INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

# ANALYSIS OF WEB USAGE MINING TECHNIQUES FOR WEB CRIME PATTERNS OF THE WEB USERS

**MS. DEEPTI VISHNU PATANGE [1], DR. P.K. BUTEY [2]**

1. CHB-Lecturer, Department of Computer Science, Arts, Science & Commerce College, Chikhaldara, Dist Amravati.
2. H.O.D., Department of Computer Science, Kamla Nehru College, Nagpur.

**Accepted Date: 05/03/2015; Published Date: 01/05/2015**

**Abstract:** Internet crime and internet security have become one of the major concerns in the modern age. There had been an enormous increase in the internet crime in the recent past. With the rapid popularity and use of the internet the instances of internet crime will also escalate. The researchers all over the world are continuously pursuing the ways and methods to curb and reduce the internet crime. In this paper we use a clustering/ classification based model to anticipate crime trends. The data mining techniques are used to analyze the web data.

**Keywords:** Web Usage Mining, Clustering, Classification, Crime Patterns

*PAPER-QR CODE*

**Corresponding Author: MS. DEEPTI VISHNU PATANGE**

**Access Online On:**

www.ijpret.com

**How to Cite This Article:**

Deepti Vishnu Patange, IJPRET, 2015; Volume 3 (9): 626-633

## INTRODUCTION

Internet crime is any illegal online activity committed on the Internet. The widespread Internet crime phenomenon encompasses multiple global levels of legislation and oversight. In the demanding and continuously changing IT field, security experts are committed to combating Internet crime through preventative technologies, such as intrusion detection networks and packet sniffers. Internet crime is a strong branch of cybercrime. Identity theft, Internet scams and cyberstalking are the primary types of Internet crime. Because Internet crimes usually engage people from various geographic areas, finding and penalizing guilty participants is complicated. Some of the very common internet crimes are as listed below.

### Types of Internet Crimes

- **Cyberbullying And Harassment**
- **Hate Crimes, Racism, Hate Websites**
- **Internet Bomb Threats**
- **Password Trafficking**
- **Identity Theft And Fraud**
- **Email Phishing**
- **Domain Name Hijacking**

Data mining has recently become one of the most progressive and promising fields for the extraction and manipulation of data to produce useful information. Thousands of businesses are using data mining applications every day in order to manipulate, identify, and extract useful information from the records stored in their databases, data repositories, and data warehouses. With this kind of information, companies have been able to improve their businesses by applying the patterns, relationships, and trends that have lain hidden or undiscovered within colossal amounts of data. For example, data mining has produced information that enables companies to create profiles of current and prospective customers to help in gaining and retaining their customers. Other uses of data mining include development of cross-selling and marketing strategies, exposure of possible crimes or frauds, finding patterns in the access of users to their web sites, and process improvement. The power of data mining is yet to be fully exploited by industry. Manufacturing, for example, is one of the new fields in which data mining tools and techniques are beginning to be used successfully. Process optimization, job shop scheduling, quality control, and human factors are some of the areas in which data mining tools such as neural networks, genetic algorithms, decision trees, and data visualization can be implemented with great results.

**RELATED WORK**

Crime Mining

Some results on crime mining have been made through using data mining techniques. Chen et al. [1] applied data mining techniques to study crime cases, which mainly concerned entity extraction, pattern clustering, classification and social network analysis. Abraham et al. [2] proposed a method to employ log files as history data to search relationship by using the frequency occurrence of incidents.

Event Oriented Construction

Event extraction is the process to extract attributes and relationship in web pages. Some researchers have proposed ideas of event oriented construction for processing events.

Lin [3] presented a method for information retrieval based on event ontology for event elements such as location, time etc. Zarri [4] proposed a method to append events for the concept of ontology to be closer to the goal of semantic web.

Focus

The focus of this research paper is on web usage mining, the focus is on the data in the web and using clustering approach. During the training phase, clustering will convert nonlinear statistical relationship between high dimensional data into simple geometrical relationship in low dimensional display.

III. METHODOLOGY

In this section we will discuss about the methodology for the research.

Data Collection

The data set is articles or documents from web pages on the internet that related to cyber terrorism. The data set consist of the text from web pages and the pictures, videos or sound format will be ignored.

Pre-processing

Pre-processing consist of the tokenization. In tokenization, all the uppercase letters are converted into lower letter words so that words can be compared and treated equally. Dictionary is used for detecting occurrence of words in the text documents.

Clustering

Then the clustering techniques are applied in order to identify the patterns in data.

Objectives of the research

To study the usage and contents pre-processing for understanding the interests of the user, filter it for crime relations.

Analytical study of the Classification and the Clustering for Pattern Discovery, to find crime patterns.

Web Usage mining involves mining the usage characteristics of the users of Web Applications as per 2.

**RESEARCH METHODOLOGY & EXPERIMENTATION**

The proposed project was implemented in 5 stages online banking system.

Procuring Data Set

The dataset of Cyber Crime Attacks for the current research work was downloaded from the website www.NSL.cs.ulb.ca/nsl/kdd.

Cleaning Data Set

A set of data items, the dataset, is a very basic concept for Data Mining. A dataset is roughly equivalent to a two-dimensional spreadsheet or database table. The dataset for crime pattern detection contained 13 attributes which were reduced to only 4 attributes by using a Java application namely sno of attack, protocol, type of attack and number of times the attack happened. This structure of 4 attributes and 50000 instances or records became the final cleaned dataset for the data mining procedures.

Processing Data Set

The data pre-processing and data mining was performed using the world famous Weka Data Mining tool. Weka is a collection of machine learning algorithms for data mining tasks. Weka is open source software for data mining under the GNU General public license. This system is developed at the University of Waikato in New Zealand. “Weka” stands for the Waikato Environment for Knowledge Analysis. Weka is freely available http://www.cs.waikato.ac.nz/ml/weka. The system is written using object oriented language

Java. Weka provides implementation of state-of-the-art data mining and machine learning algorithm. User can perform association, filtering, classification, clustering, visualization, regression etc. by using Weka tool.

Each and every organization is accession vast and amplifying amounts of data in different formats and different databases at different platforms. This data provides any meaningful information that can be used to know anything about any object. Information is nothing just data with some meaning or processed data. Information is then converted to knowledge to use with KDD.

**PREFORMING ANALYSIS ON DATA SET**

There are two methods used in the current study for generating results as below

K-Means Algorithm

The K-means [5] algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into K groups, where K is provided as an input parameter. It then assigns each observation to clusters based upon the observation proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again. The working of algorithm is explained as follows:

The algorithm arbitrarily selects K points as the initial cluster centres.

Each point in the data set is assigned to the closed cluster based on the Euclidean distance between each point and each cluster centre.

Each cluster centre is recomputed as the average of the points in that cluster.

Steps 2 and 3 repeats until the clusters converge.

Convergence may be defined differently depending upon the implementation, but it normally means that either no observation change cluster when step 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters [7].

The k-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters [8].

## RESULTS AND ANALYSIS

The dataset was preprocessed using Weka tool for all 4 attributes and statistical output was produced by Weka with respect to Minimum, Maximum value, Mean and the Standard Deviation. Further, the zero attributed was removed from the data set and the data set was again preprocessed using Attribute Selection in the Weka software. This step depicted the count as number of attacks and the weighted average of the attacks made through 3 different protocols namely TCP, UDP and ICMP as shown in fig 1.

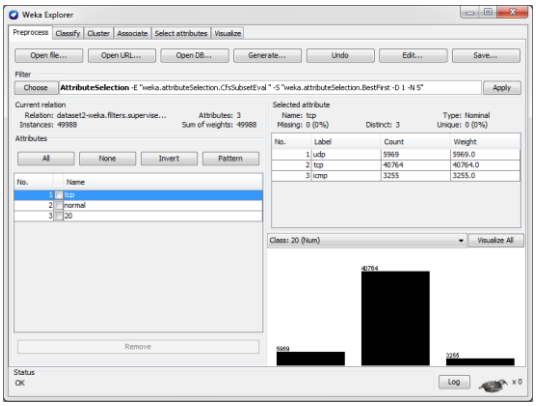

**Fig. 1**

The frequency of the attacks based on the network protocol was generated using visualization method in the Weka tool. The type of attacks was represented using different colors. Fig 2 show the graphical representation of the attack frequency.

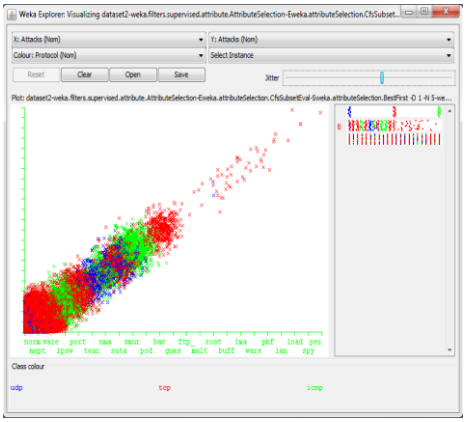

**Fig. 2: Graphical Attack Frequency**

The classification was made on the basis of Protocols and the results were obtained from the Weka program. The classification was made on the 49988 instances of the dataset with 3 attributes. The visualization of Error graph on protocols after classification is shown in fig 3. and the visualization of Error graph on attacks

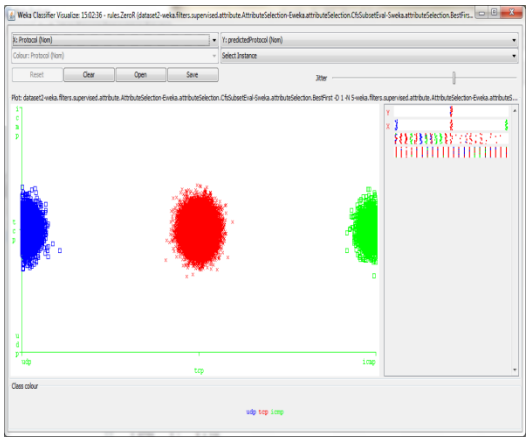

**Fig.3**

The visualization of Error graph on attacks after classification is depicted in fig 4.

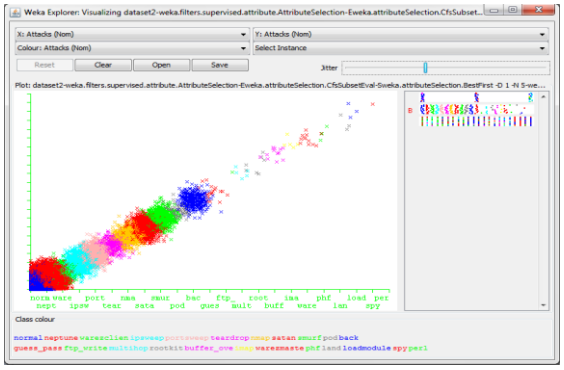

**Fig.4**

## CONCLUSION

Internet crime detection is significant in the modern age. Despite the facts like the extensive and rapid growth of internet, very massive online financial and commercial activities, storage of volumes of users private data, there is lack of truly secured system which makes it an important field of research. An effective web crime detection system should be able to discover both the known and new attacks as soon as possible. This research work uses an algorithm for constructing patterns of data using clustering algorithm. Firstly, the K-means clustering

algorithm is used to obtain results to predict the malicious attacks behavior of the users implementation incorporated in the WEKA data mining tool. From fig. 4 maximum attacks were performed by Neptune attack among all since users are more inclined to use tcp protocol.

**REFERENCES**

1. Hsinchun Chen, Wingyan Chung, Yi Qin, et al. Crime Data Mining: An Overview and Case Studies. Proceeding of the 2003 annual national conference on Digital government research, Boston, M.A, 2003, pp 1-5.

2. T. Abraham and O. de Vel. Investigating profiling with computer forensic log data and association rules. Proc. Of the IEEE International Conference on Data Mining (ICDM'06), 2006, pp 11-18.

3. H. F. Lin and J. M. Liang Event based ontology design for retrieving digital achieves on human religious self-help consulting. Proc. Of 2005 IEEE International Conference on e-technology, e-Commerce and e-Service, 2005 pp. 453-475.

4. G. P. Zarri. Semantic web and Knowledge Representation, Proc. Of the 13th International Workshop on Database and Expert System Applications (DEXA'02), 2002, pp. 1529-4188.

5. Teknomo, Kardi, "K-means Clustering Tutorials".

6. Malathi. A, Dr. S. Santosh Baboo and Anbarasi. An intelligent analysis of city crime data using data mining. International Conference on Information and Electronics Engineering IPCSIT Vol 06. pp. 130-134.

7. http://databases.about.com/od/datamining/a/kmeans.htm.

8. Sheilini Jindal , "A Proportional Analysis On The Illustrious Practices For The Extraction And Discovery Of Hidden Patterns - Data And Web Mining", International Journal of Enterprise Computing and Business Systems (Online)http://www.ijecbs.com, Vol. 1 Issue 1 January 2011.