ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 2, Issue 3, September 2012

An Intelligent Analysis of Web Crime Data Using Data Mining

Anshu Sharma, Shilpa Sharma

***Abstract*—there had been an enormous increase in the crime in the recent past. The concern about national security has increased significantly. With the rapid popularity of the internet, crime information on the web is becoming increasingly rampant. A lot of crime information in documents is described through events. In this paper we construct the scenario to extract the attributes and relations in the web pages and reconstruct the scenario for crime mining. 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.**

Index Terms—**Web Mining, Cyber Crime, Classification, Clustering.**

I. INTRODUCTION

Web crime is the use of internet based attacks in crime activities, including acts of deliberate, large scale disruption of computer network, especially of the personal computer attach to the internet by the means of tools such as viruses. Web crime can be basically divided into two major activities. One is those take the network as criminal objects and the other are those using the network to commit crime such as fraud, eroticism etc. One of the challenges to web crime is the difficulty of analyzing large volumes of data involved in criminal and terrorist activities [6]. Data mining makes it easy, convenient and practical to explore very large databases for organizations and users. A web page involving a crime can be thought of as a chain of actions with series of background attributes. We can analyze web information from the perspective of events and apply some research results related to the events to solve the problem of web crime mining.

II. RELATED WORK

A. *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.

B. *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.

C. *Focus*

The focus of this research paper is on web content mining, the focus is on the text in the web and using clustering approach. During the training phase, clustering will convert non linear 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.

A. *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.

B. *Preprocessing*

Preprocessing 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.

C. *Clustering*

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

IV. EVENT ORIENTED WEB CRIME CONSTRUCTION

A. *Categories of Cyber Crimes*

The internet as tools can be divided into the following categories: (1) fraud; (2) internet pornography; (3) illegal trade; (5) false advertising; (6) violations of privacy; (7) forgery; (8) network gambling etc.

B. *Events and Their Relations*

An event is defined as the change in system state. It refers to one thing happening at a specific time and location. In case of supervised techniques one will analyze each event to determine how similar it is to the majority and their success depends on the choice of similarity measures and dimension weighting, while in case of unsupervised techniques one will build a model for rare events based on labeled data and use it to classify each event.

V. WEB CRIME MINING BASED EVENT CONSTRUCTION

A. *Process*

The process includes the following modules:

- **Pre-processing web texts:** the web texts are generally in the HTML format. Pre-processing is applied to filter the details and getting the pure texts.
- **Candidate features:** after applying the filtering process the remaining words left are the candidate features of the pure texts.
- **Feature reconstruction:** in this step the various features are merged and some features are appended to obtain the desired result.
- **Data Mining:** various data mining techniques such as classification, clustering etc are applied on the web text to obtain the patterns of data.

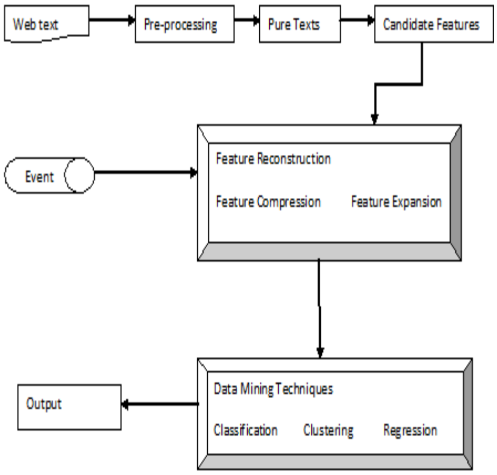


Fig1: Process of Web Crime Mining

The above figure presents the overall process of web crime mining.

B. *Feature Reconstruction*

The features of a document are high dimensional and sparse. There are two main techniques applied to a document. One is feature compression and the other is feature expansion. Feature compression improves the precision of the text processing. Feature compression is used to compress feature dimension through merging some features. Many synonyms that are used in a document are merged using the concept of feature compression. Feature expansion technique is used to process by expanding the features. The documents are expanded by adding the events in order to complete the omitted elements from the document. Feature expansion can also be used in case of short documents by placing the words related to the document in the present document.

C. *Web Data Mining Techniques*

Traditional data mining techniques such as association analysis, classification, clustering analysis are used to identify patterns in structured data. There are newer techniques that can be used to identify patterns in structured as well as in unstructured data. Entity extraction identifies particular patterns of data such as text, images and audio data etc. entity extraction provides basic information for crime analysis, but its performance depends greatly upon the availability of excessive amount of clean input data. Clustering techniques group data items into classes with similar characteristics to maximize or minimize similarity. Clustering crime incidents can automate major part of crime analysis but it is limited by the high computational intensity typically required.

D. *K-Means Clustering 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:

1. The algorithm arbitrarily selects K points as the initial cluster centers.
2. Each point in the data set is assigned to the closed cluster based on the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. 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].

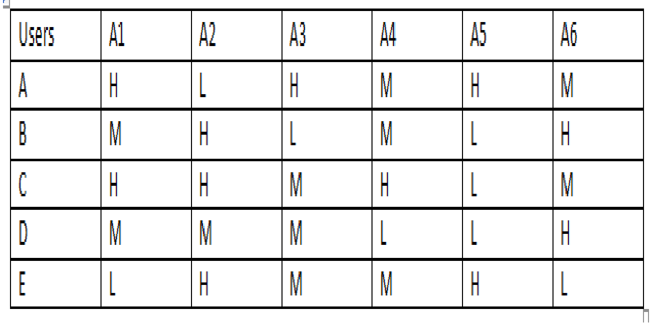
VI. IMPLEMENTATION DETAILS

Experiments are conducted on real world data of few users as well as on the synthetically generated data of various users with different kind of usage behavior. The collected data is distributed into three categories, two third of genuine data GD and fraud history data FD are used for training and remaining datasets are used for prediction and web crime detection. The various implementation steps are:

A. Collect the User Data

Companies are not ready to share information of their employees. Therefore the performance of the proposed system has been tested on five real users' data and on synthetically generated data by the simulator. For describing the implementation steps we select three legal and two crime data sets which are given in table I.

Table I: Sample Data



Users	A1	A2	A3	A4	A5	A6
A	H	L	H	M	H	M
B	M	H	L	M	L	H
C	H	H	M	H	L	M
D	M	M	M	L	L	H
E	L	H	M	M	H	L

B. Select the Training Sample

The profile size can be chosen appropriately. In the case of existing crime detection large profile size increases the accuracy but at the same time it increases the training time.

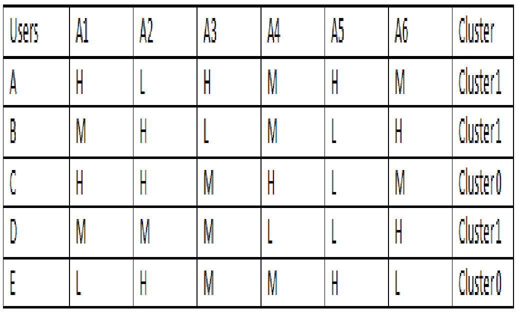
C. Data Cleaning

Select only the required attributes and discard the others. Sample transaction data used for training after cleaning is given in table I.

D. Data Transformation

Classify the data into two clusters- cluster 0(genuine user) and cluster I (illegal data detected) using K-means clustering algorithm. This work selects K as two. After executing K-means algorithm on the sample transaction in table I, we get the cluster which is given in Table II.

Table II: Result of Clustering



Users	A1	A2	A3	A4	A5	A6	Cluster
A	H	L	H	M	H	M	Cluster1
B	M	H	L	M	L	H	Cluster1
C	H	H	M	H	L	M	Cluster0
D	M	M	M	L	L	H	Cluster1
E	L	H	M	M	H	L	Cluster0

From the above observation it is clear that users C and E are in cluster 0 and users A, B and D are in cluster 1. Now the users in cluster 1 are under further investigation to reduce the false alarm rate. The false alarm rate refers to the check which is performed to verify whether any illegal user fall under genuine user case.

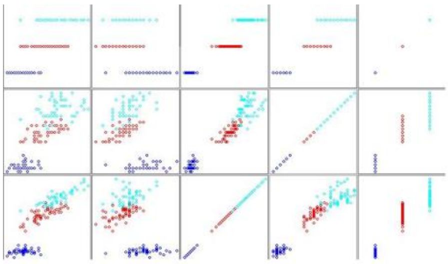


Fig 2: Cluster Distribution Plot

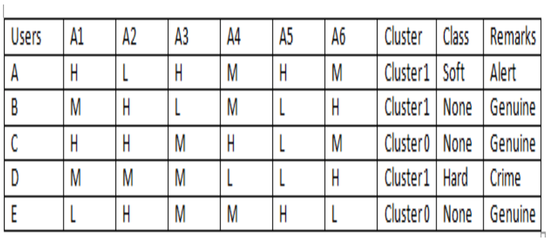
E. Training Model Using Classification

Now when the complete data is clustered into two groups and in order to reduce the false alarm rate and for the detection of web crime we will pay attention to the datasets belonging to the cluster 1. We train the model using classification technique.

VII. RESULTS

In every model, the accuracy plays an important role in the acceptance of that model for the application. Table III shows main results of the implementation of the user data described in table I. the accuracy of the clustering comes out to be 94.75% and 5.28% comes under false alarm rate.

TABLE III: FINAL OUTPUT



Users	A1	A2	A3	A4	A5	A6	Cluster	Class	Remarks
A	H	L	H	M	H	M	Cluster1	Soft	Alert
B	M	H	L	M	L	H	Cluster1	None	Genuine
C	H	H	M	H	L	M	Cluster0	None	Genuine
D	M	M	M	L	L	H	Cluster1	Hard	Crime
E	L	H	M	M	H	L	Cluster0	None	Genuine

VIII. CONCLUSION

Web crime detection is important in today's internet environment. The combination of facts such as extensive growth of internet, the vast financial possibilities and the lack of truly secured system 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 uses the scalable algorithm for constructing patterns of data using clustering algorithm. K-means clustering algorithm is used and then data is classified to obtain crime, none and genuine users. The accuracy of the proposed work comes out to be 94.75% and it efficiently detects the false rate anomalies.

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 archieves 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.

AUTHOR'S PROFILE

Anshu Sharma received her B.TECH degree From Punjab Technical University in 2009. She is currently doing M.TECH from Punjab Technical University. She has 3 years of teaching experience. Her Research Interests include Data Mining, Software Engineering, Expert Systems, and Real Time Systems etc.

Shilpa Sharma has done BTECH from PTU and currently pursuing her MTECH from PTU. She works as a Lecturer in Lovely Professional University. She has 3 years of Experience as a Lecturer. Her Research Interests Include Data Mining, Software Engineering, Expert Systems and Neural networks Etc.