ISSN:0975-9646

Unauthorized Terror Attack Tracking Using Web Usage Mining

Ramesh Yevale, Mayuri Dhage, Tejali Nalawade,.Trupti Kaule.

Dept. Of computer Science & Engineering
Shriram Institute Of Engineering & Technology Centre
Paniv Tal-Malshiras Dist-Solapur

Abstract -**Terrorist groups use the Web as their infrastructure for various purposes. One example is the forming of new local cells that may later become active and perform acts of terror. The Terror Tracking using Web Usage Mining (TTUM) is aimed at tracking down online access to abnormal content, which may include terrorist-generated sites, by analyzing the content of information accessed by the Web users. TTUM operates in two modes: the training mode and the detection mode. In the training mode, TTUM determines the typical interests of a prespecified group of users by processing the Web pages accessed by these users over time. In the detection mode, TTUM performs real-time monitoring of the Web traffic generated by the monitored group, analyzes the content of the accessed Web pages, and issues an alarm if the accessed information is not within the typical interests of that group and similar to the terrorist interests. An experimental version of TTUM was implemented and evaluated in a local network environment. An innovative knowledge-based methodology for terrorist tracking by using Web traffic content as the audit information is presented. The proposed methodology learns the typical behavior ('profile') of terrorists by applying a data mining algorithm to the textual content of terror-related Web sites. The resulting profile is used by the system to perform real-time detection of users suspected of being engaged in terrorist activities. The Receiver-Operator Characteristic (ROC) analysis shows that this methodology can outperform a command based intrusion detection system.**

Keywords: Data mining System architectures, Data mining application , Dataset, Crime, Terrorism, Warehouse, Knowledge Discovery Database.

1. TERROR TRACKING WEB MINING

Terrorism is defined as: "An illegal preconceived use of physical or psychic violence (or threat of it) for further political goals aimed at civilians or non-combatants etc., to change current policies, their methods and structure". Web Mining is the application of data mining and information extraction techniques aimed at discovering Patterns and knowledge from the Web. Traditionally, Web Mining is divided into three classes:

- Web Content Mining - discovery of useful information from text, image, audio or video data in the Web.
- Web Structure Mining - analysis of the node and connection (graph) structure underlying single web sites, as well as larger collections of interrelated sites
- Web Usage Mining - often called Web analytics involves extracting useful information from server logs and other sources detailing usage patterns.

A. Web content mining -

Web content mining is related but different from data mining and text mining. It is related to data mining because many data mining techniques can be applied in Web content mining. It is related to text mining because much of the web contents are texts. However, it is also quite different from data mining because Web data are mainly semi-structured and/or unstructured, while data mining deals primarily with structured data. Web content mining is also different from text mining because of the semi-structure nature of the Web, while text mining focuses on unstructured texts. Web content mining thus requires creative applications of data mining and/or text mining techniques and also its own unique approaches. In the past few years, there was a rapid expansion of activities in the Web content mining area. This is not surpnsmg because of the phenomenal growth of the Web contents and significant economic benefit of such mining.

B. Web Structure Mining –

World Wide Web can reveal more information than just the information contained in documents. For example, links pointing to a document indicate the popularity of the document, while links coming out of a document indicate the richness or perhaps the variety of topics covered in the document. This can be compared to bibliographical citations. When a paper is cited often, it ought to be important. By means of counters, higher levels cumulate the number of artifacts subsumed by the concepts they hold. Counters of hyperlinks, in and out documents, retrace the structure of the web artifacts summarized.

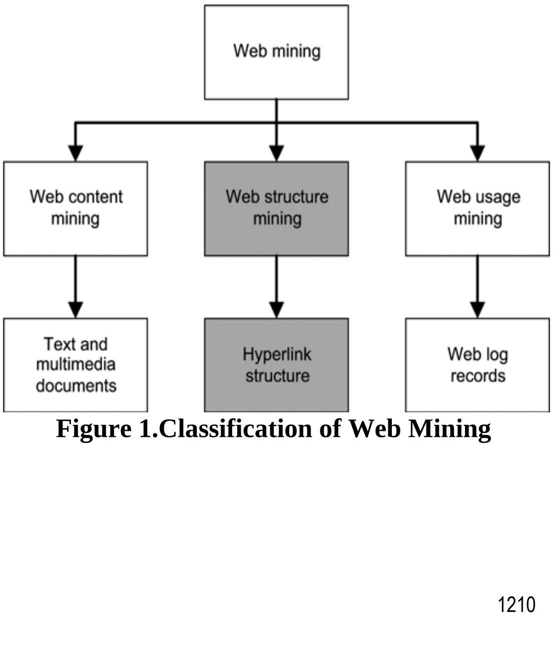


Figure 1.Classification of Web Mining

1.1. Web Usage Mining System Structure

Web Usage Mining-Web Usage Mining is a part of Web Mining, which, in turn, is a part of Data Mining. As Data Mining involves the concept of extraction meaningful and valuable information from large volume of data, Web Usage mining involves mining the usage characteristics of the users of Web Applications. This extracted information can then be used in a variety of ways such as, improvement of the application, checking of fraudulent elements etc. Web Usage Mining is often regarded as a part of the Business Intelligence in an organization rather than the technical aspect. It is used for deciding business strategies through the efficient use of Web Applications.

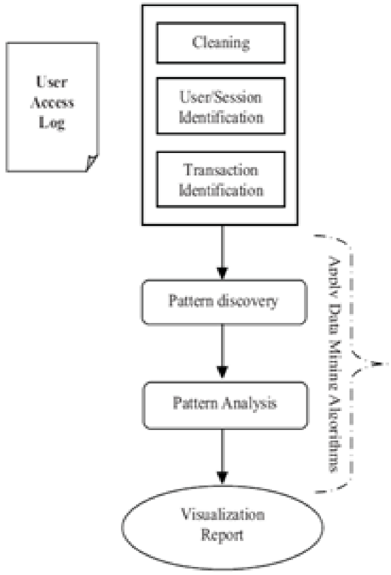

User Access Log
Cleaning
User/Session Identification
Transaction Identification
Pattern discovery
Pattern Analysis
Visualization Report
Apply Data Mining Algorithms

Figure 2. Web Usage Mining System Structures

SCOPE OF RESEARCH WORK

The purpose of the project is that the project is very useful to find out the terror activities. By this project, users of the project can come to know across the suspicious conversation, and thus can detect the source of the terror and will be useful to minimize the terror activities. The project has very wide scope in the security of the national assets. The project is very useful to find out terror activities. It is mainly useful for the Military and CBI agents. By using this they can come to know the different terror communication and thus the accidents can be avoided. Blogs, social networking sites, mostly hosted by terrorist sympathizers is avoided by using web usage mining.

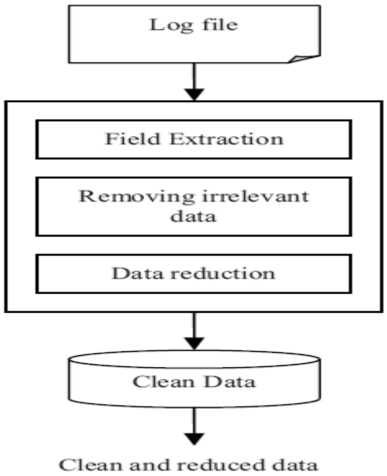
LITERATURE SURVEY:

Basis of study idea:

Web mining is a rapidly growing research area. It consists of Web usage mining, Web structure mining, and Web content mining. Web usage mining refers to the discovery of user access patterns from Web usage logs. Web structure mining tries to discover useful knowledge from the structure of hyperlinks.

PROPOSED SYSTEM

Working System flow



A. Data Storage: The results of preprocessing the web server logs are stored in a relational database to facilitate easy retrieval and analysis.

Data preprocessing: Preprocessing converts the raw data into the data abstractions necessary for pattern discovery. The purpose of data preprocessing is to improve data quality and increase mining accuracy. Preprocessing consists of field extraction, data cleaning. This phase is probably the most complex and ungrateful step of the overall process. This system only describe it shortly and say that its main task is to ”clean” the raw web log files and insert the processed data into a relational database, in order to make it appropriate to apply the data mining techniques in the second phase of the process. So the main steps of this phase are:

1) Extract the web logs that collect the data in the web server.

2) Clean the web logs and remove the redundant information.

3) Parse the data and put it in a relational database or a data warehouse and data is reduced to be used in frequency analysis to create summary reports.

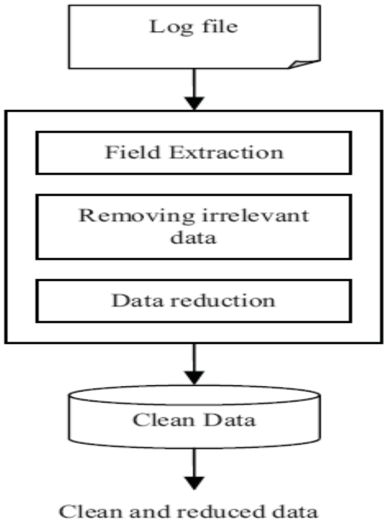


Figure 3. Web log data pre-processing

B. **Field Extraction:** The log entry contains various fields which need to be separate out for the processing. The process of separating field from the single line of the log file is known as field extraction. The server used different characters which work as separators. The most used separator character is or 'space' character.

C. Data Cleaning

Data cleaning eliminates irrelevant or unnecessary items in the analyzed data. A web site can be accessed by millions of users. The records with failed HTTP status codes also may involve in log data. Data cleaning is usually site-specific, and involves extraneous references to embedded objects that may not be important for purpose of analysis, including references to style files, graphics or sound files. Therefore some of entries are useless for analysis process that is cleaned from the log files. By Data cleaning, errors and inconsistencies will be detected and removed to improve the quality of data. Algorithm had not only cleaned noisy data but also reduced incomplete, inconsistent and irrelevant requests according to step 4 and 5. Error requests are useless for the process of mining. These requests can be removed by checking the status of request. For example, if the status is 404, it is shown that the requested resource is not existence. Then, this log entry in log files is removed. Moreover, unnecessary log data is also eliminated URL name suffix, such as gif, jpg and so on in step **6** and **7**. Finally, usefulness and consistent records remain in SLT of database after data cleaning.

Advantages

- It provides marketing intelligence.
- Web logs provide an exciting new way of collecting information on visitors.
- To create personalized search engines, which can understand a person's search queries in a personal way by analyzing and profiling the user's search behavior?
- The data mining methods used to accurately detect malicious executables before they run.
- Integrate data sources Clean/ modify data sources Build Profiles of Terrorists and Activities Examine results/ Prune results Report final results Data sources with information about terrorists and terrorist activities mine the data.

RELATED WORKS

Algorithm used

1. FIELD EXTRACTION ALGORITHM

The Field Extraction algorithm is given below.

```
Input: Log File
Output: DB
Begin
Step 1:  Open a DB connection
Step 2:  Create a table to store log data
Step 3:  Open Log File
Step 4:  Read all fields contain in Log File
Step 5:  Separate out the Attribute in the string Log
Step 6:  Extract all fields and Add into the Log Table (LT)
Step 7:  Close a DB connection and Log File
End
```

2. DATA CHAINING ALGORITHM

```
Input: Log Table (LT)
Output: Summarized Log Table (SLT)
'*' = access pages consist of embedded objects
(i.e. .jpg, .gif, etc)
'**' =successful status codes and requested methods (i.e. 200, GET etc)
Begin
Step 1:  Read records in LT
Step 2:  For each record in LT
Step 3:  Read fields (Status code, method)
Step 4: If Status code='**'and method= '**'
Then,
Step 5: Get IP address and URL link
Step 6: If suffix. URL Link= {*.gif,*.jpg,*.css}
Then
Step 7: Remove suffix. URL link
Step 8: Save and URL Link
End if
Else
Step 9:  Next record
End if
End
```

CONCLUSION

In this paper we are discuss the Data pre-processing is an important task of TTUM application. Thus web mining technique can be used for detecting and avoiding terror threats caused by terrorists all over the world. Data mining and web data mining technologies will have a significant impact on counter-terrorism. As we are seeing, one of the major concerns of our nation today is to detect and prevent terrorist attacks. This is also becoming the goal of many nations in the world. We need to examine the various data mining and web mining technologies and see how they can be adapted for counter-terrorism.

REFERENCES

[1] •Theint Theint Aye ,"Web Log Cleaning for Mining of Web Usage Patterns" 2011 IEEE.

[2] •Shaily Langhnoja,Mehul Barot, Darshak Mehta," Pre-Processing: Procedure on Web Log File for Web Usage Mining" International Journal of Emerging Technology and Advanced Engineering December 2012.

[3] •4.L.K.JoshilaGrace,V.Maheswari,Dhinaharan Nagamalai," Analysis of Web Logs and Web User In Web Mining" International Journal of Network Security & Its Applications (IJNSA),Vol.3,No.1,January 2011.

[4] •J.Vellingiri,S. Chenthur Pandian,"A Survey on Web Usage Mining" Volume 11 Issue 4 Version 1.0 March 2011.

[5] •Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremistgroup Web forum messages. IEEE Intelligent Systems, Special Issue on Artificial Intelligence for National and Homeland Security, 20(5),[67–75].