International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-2 Issue-5, April 2013

# Web Usage Mining Based on Complex Structure of XML for Web IDS

**Marjan Eshaghi, S.Z. Gawali**

***Abstract- In current trend, most of the businesses are running through online web applications such as banking, shopping, and several other e-commerce applications. Hence, securing the web sites is becomes must do task in order to secure sensitive information of end users as well as organizations. Web log files are generated for each user whenever he/she navigates through such e-commerce websites, users every click is recorded into such web log files. The analysis of such web log files now a day's done using concepts of data mining. Further results of this data mining techniques are used in many applications. Most important use of such mining of web logs is in web intrusion detection. To improve the efficiency of intrusion detection on web, we must have efficient web mining technique which will process web log files. In this project, our first aim is to present the efficient web mining technique, in which we will present how various web log files in different format will combined together in one XML format to further mine and detect web attacks. And because log files usually contain noisy and ambiguous data this project will show how data will be preprocessed before applying mining process in order to detect attacks. Hence mining process includes two parts, web log files preprocessing in order to remove the noise or ambiguous data mining process to detect the web attacks.***

***Keywords— log files, web mining, preprocessing, IDS, XML, CRM.***

## I. INTRODUCTION

Web Usage Mining is commonly considered a region of the Business Intelligence in a corporation instead of the technical facet. It's used for deciding business ways through the economical use of net Applications. It's additionally crucial for the client Relationship Management (CRM) because it will guarantee client satisfaction as so much because the interaction between the client and also the organization worries.

The major drawback with net mining generally and net Usage Mining above all is that the nature of the info they deal with. With the upsurge of web during this millennium, the online knowledge has become vast in nature and lots of transactions and usages are happening by the seconds. With the exception of the amount of the info, the info isn't fully structured. It's in a very semi-structured format so it wants lots of preprocessing and parsing before the particular extraction of the desired info.

**Manuscript published on 30 April 2013.**
*Correspondence Author(s)
**Marjan Eshaghi,** Information Technology Department, College of Engineering, Bharati Vidyapeeth University,Pune, India.
**Prof S.Z. Gawali,** Information Technology Department, College of Engineering, Bharati Vidyapeeth University,Pune, India.



As we studied in [1], author presented the web mining based on web log files, he suggested the data preprocessing and its importance, however in algorithm author only concentrate on mining of XML files using the clustering approach. Another problem associated with this approach is that, the proposed method of web mining is not yet tested over the real time environment or any of applications like intrusion detection, users behavior prediction etc. In addition to this, the architecture given in [1] shows that we have to give set of web log files as input which may of different types. However log files with different format needs to be combined in order to form the XML file; these things are not clearly presented.

The problem in the existing approach given in the [1] is that preprocessing methodology and log files various formats combining methodology not so efficient. In the current era, we are witnessing a surge of Web Usage around the globe. A large volume of data is constantly being accessed and shared among a varied type of users; both humans and intelligent machines. Thus, taking up a structured approach to control this information exchange, has what made Web Mining one of the hot topics in the field of Information Technology. In addition to this, increasing web based attacks now days also one of the key factor behind this project, we feel that if we have efficient preprocessing and web mining concepts in place we can detect the attacks on web efficiently.

## II. EXISTING WORK AND LIMITATIONS

There are methods used in order to prevent SQL attacks, and one of them is the use of Proxy Filters, which is a system of enforcing input validation rules on data that are flowing the to a web application. The developers offer constraints through the Security Policy Descriptor Language (SPDL), thus specifying the transformations that are applied for application of parameters that flowing Web page to the application server [11]. This method also allows developers to express their policies since SPDL is highly expressive, though the approach is human-based and defensive programming, thus requiring the developers to identify the data that require filtering.

## III. SCOPE OF RESEARCH

As we know that Web Usage Mining important parts of web mining nowadays, which, in turn, a part of data processing. As data processing involves the idea of extraction significant and valuable info from massive volume of knowledge, internet Usage mining involves mining the usage characteristics of the users of internet Applications.

This extracted info will then be employed in a range of the way like, improvement of the appliance, checking of dishonest components, and detection of attacks. Scope of this project is present in the efficient web logs preprocessing, web mining and extraction of web attacks.

## IV. SYSTEM ANALYSIS AND DESIGN

### *System Analysis*

Web server log are files store user click streams whereas navigating an internet website. a number of these knowledge area unit supernumerary for the analysis method and will have an effect on the detection of net attacks. Therefore, preprocessing step comes before applying mining algorithms. Sadly, most of the researches during this topic provide no details concerning preprocessing steps.

### *System Description*

We have drawn new approach in preprocessing of diary files for internet intrusion detection. We have a tendency to mention the various steps during this method and therefore the variations in these steps from internet usage mining. Additionally, we have a tendency to illustrate a way to mix log files with completely different formats in one common place format victimization XML. We have a tendency to provide to algorithms to mix those log files. These algorithms are enforced victimization c# code and underneath windows visual image package.

## V. PROPOSED ALGORITHM

They just mention implicitly that these log files ought to be reborn into appropriate format. During this paper we are going to discuss this issue. Figure illustrates steps concerned in preprocessing method. It involves desegregation information from multiple log files into one file with one format. The subsequent subsections can contain details regarding these steps.

A web server log file may be a straightforward plain computer file that records info anytime a user requests a resource from an online website. Access log file, Error log file, Agent log file, Referrer log file. Figure 1 showing the proposed approach of this paper:

### *Data Integration*

Sometimes one data processor is hosted on totally different completely different servers; this can produce different log files for identical data processor. Desegregation these log files along makes them a lot of valuable and allows a lot of data extraction. The format of those log files depends on the configuration of the online server. Here we'll take into account to formats: W3C format and NCSA common format. Log files are often combined either in text format or regenerate to electronic information service.

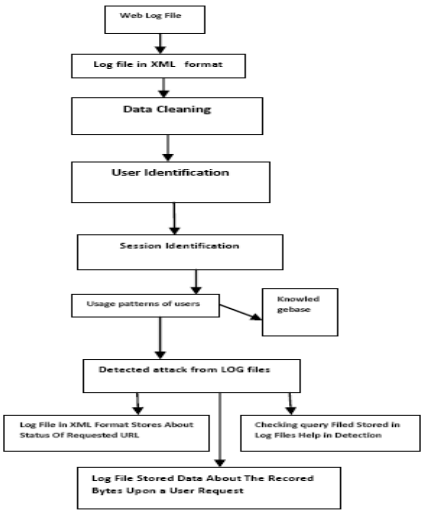



Figure 1: Proposed Approach

However we'll mix them in XML files. XML files area unit a lot of structured and a lot of decipherable than text format and that they area unit easier and need less cupboard space than electronic information service. The tags accustomed record entries of those to files are: user-IP, user-name, date, time, uri-stream, uri-query, status, bytes-sent. But date and time field's area unit recorded in several format within to log files.

### *Data Cleansing*

Data cleansing method is to get rid of abuzz and supererogatory knowledge that will have an effect on the mining method. The input for this step is that the XML file that contains the combined log files. The info cleansing includes the subsequent steps:

1) Take away log Entry nodes that contain in uri-stem kid node extensions like jpg, gif, css. This step is common with cleansing method in internet usage mining

2) In contrast to preprocessing internet usage mining, standing with code four hundred series and five hundred series ought to be unbroken as a result of they will be thought of as anomaly actions.

3) Take away log Entry nodes with undefeated standing (200 series) and with "-"in uri-query node. As a result of likelihood of requests with such characteristics to contain internet attacks is nearly zero. Though question string related to address requests could embody internet attacks like XSS or SQL injection, these address could execute with success if these attacks don't seem to be detected. This can be why any parameters ought to be examined notwithstanding its standing is 200 series.

### *User Identification*

In the e-commerce context not like alternative internet primarily based domains user identification may be a easy drawback as in most cases customers should login victimisation their distinctive ID. We tend to have an interest in anonymous users just in case they cause lots of errors or their uri-query node isn't empty. It's not vital to spot the identity of this user, what's vital is to discover attack he she triggers if there.


Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

***Session Identification***

For best-known users, i.e. users logged in e-commerce sites with user name, session identification is time directed. If the time between requests exceeds half-hour, this suggests the beginning of recent session[12] [5]. An anonymous user, what's necessary is to sight that IPs causes error in tiny amount of your time.

Unlike in net usage mining, when the preprocessing method no ought to convert knowledge in XML file to electronic information service. We will use XML file as input to the mining method for the detection of net attacks.

***Detected attacks from log files***

In the previous section we have a tendency to known however log files is preprocessed to be prepared for detection of intrusions. During this section we are going to illustrate however the previous preparations of log files facilitate in detection of various sorts of intrusion.

The log go in the format of XML stores field concerning the standing of the requested computer address, this will facilitate in:

1) Brute force attack: the user that triggers plenty of errors in little amount of your time is suspicious to brute force attack. This will be detected through checking of the standing field.

2) Checking standing field facilitate in distinctive malicious users that trigger plenty of errors whereas browsing the positioning

Checking the uri-query field hold on in log files facilitate in detection of:

1) SQL injection, X Path injection, XSS attacks: these attacks are detected from checking the uri-query field for dangerous keywords which will reveal the prevalence of those attacks. This checking is important albeit the standing field doesn't contain error series code.

2) The modification in parameters order or the absence of some needed parameters for constant page could indicate anomaly actions. Log files additionally store knowledge concerning the came back bytes upon a user request this helps in:

Anomaly behavior is detected once user requests a page and it's come back bytes are completely different from alternative requests for constant page.

## VI. EXPERIMENTAL ANALYSIS

***Requirement Analysis***

Microsoft .NET profiling API is a standard mechanism for application or tool developers to instrument the behavior of a .NET program by providing components that register events of interest, such as just-in-time compilation or method invocation. Right before a .NET managed method is invoked, Common Language Runtime (CLR) loads the .NET assembly that contains the method code into memory, and compiles it into executable by the target CPU (a process called just-in-time compilation).

ASP.NET version of SQL Interceptor provides components that register just-in-time compilation events of database-access classes such as SQLCommand and OleDbCommand. In the event handler,SQL Interceptor employs MSIL re-write technique to install a hook that calls SQLIA Detector module in the beginning of each method of interest. Those mechanisms are supported by all versions of .NET, and can be deployed into existing ASP.NET web application dynamically without access to application source code. ASP, executed in the domain of COM (Common Object Model), has no default mechanism for intercepting SQL statements, either.

***Hardware and Software requirements***

HARDWARE REQUIREMENTS

- Windows XP
- RAM – 1GB
- Hard Disk - 40GB

SOFTWARE REQUIREMENTS

- .Net Framework
- IIS 7.0
- MySQL
- SQL Server

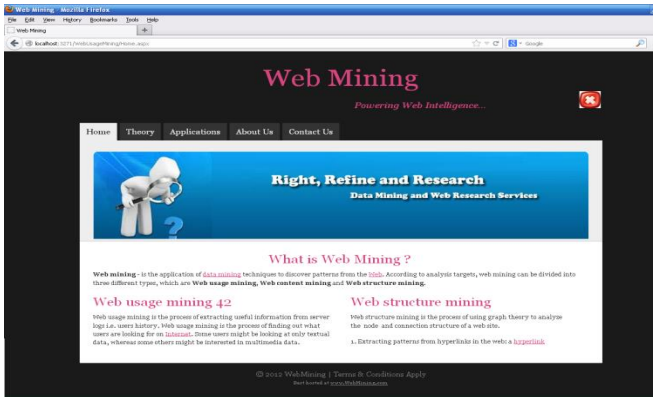
***Applications***

**1. Letizia:** Letizia is an application that assists a user browsing the Internet. As the user operates a conventional Web browser such as Mozilla, the application tracks usage patterns and attempts to predict items of interest by performing concurrent and autonomous exploration of links from the user's current position. The application uses a best-first search augmented by heuristics inferring user interest from browsing behavior.

2. WebSift: The WebSIFT (Web Site Information Filter) system is another application which performs Web Usage Mining from server logs recorded in the extended NSCA format (includes referrer and agent fields), which is quite similar to the combined log format which used in case of DSpace log files. The preprocessing algorithms include identifying users, server sessions, and identifying cached page references through the use of the referrer field. It identifies interesting information and frequent item sets from mining usage data.

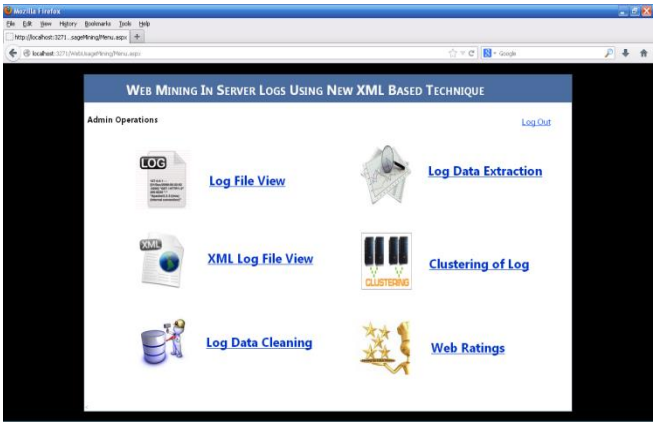
3. Adaptive Websites: An adaptive website adjusts the structure, content, or presentation of information in response to measured user interaction with the site, with the objective of optimizing future user interactions. Adaptive websites are web sites that automatically improve their organization and presentation by learning from their user access patterns. User interaction patterns may be collected directly on the website or may be mined from Web server logs.

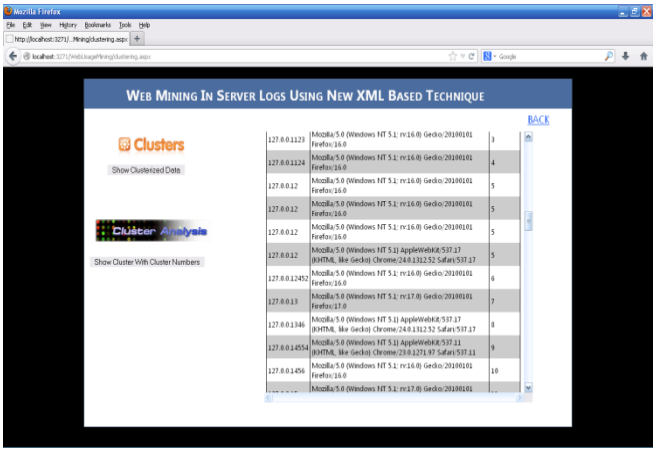
5.4 Results

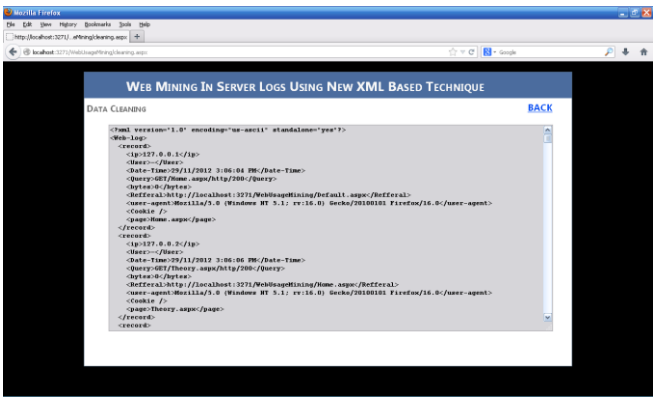




Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)




## VII. CONCLUSION AND FUTURE WORK

SQL injection vulnerabilities are omnipresent and dangerous; nevertheless several web applications as we discussed above, in this project we presented new approach in preprocessing of web log files for web intrusion detection. We discussed the different steps in this process and the differences in these steps from web usage mining. In addition, we illustrated how to combine two log files with different formats in one standard format using XML. We provided two algorithms to combine those log files. These algorithms have been implemented using c# code and under windows vista operating system.

## REFERENCES


1. "XML Based Web Usage Mining In Server Logs", Y.S.S.R Murthy, L.Balaji & Lakshmi Tulasi.Ambat.
2. A. Hamami, M. Ala'a, S. Hasan. (2006). Applying Data Mining Techniques in Intrusion Detection System on Web and Analysis of Web Usage, Information Technology Journal, 2006.
3. C.J. Ezeife, J. Dong, A.K. Aggarwal. (2007). SensorWebIDS: A Web Mining Intrusion Detection System, International Journal of Web Information Systems, volume 4, pp. 97-120, 2007.
4. C. Kruegel, G. Vigna. (2003). Anomaly Detection of Web-based Attacks, CCS, 2003.
5. G. Shiva, N.V. Suba, U. Dinesh. (2010). Knowledge Discovery from Web Usage Data: A survey of Web Usage Pre-processing Techniques, Springer, 2010.
6. Andrews, M.: Guest Editor's Introduction: The State of Web Security. IEEE Security and Privacy, 4, 4, 14--15 (2006)
7. K.R. Suneetha, Dr. R. Krihnamoorthi. (2009). Identifying User Behavior by Analyzing Web Server Access Log File, IJCSNS, 2009.
8. L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai. (2011). Analysis of web logs and web user in web mining, IJNSA, 2011.
9. Jaideep Srivastava , Robert Cooley, Mukund Deshpande, Pang-Ning Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations, Volume 1, Issue 2-Pages 12-23.
10. Adel T. Rahmani and B. Hoda Helmi, EIN-WUM an AIS-based Algorithm for Web Usage Mining, Proceedings of GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA, ACM978-1-60558-130-9/08/07 (Pages 291-292)
11. Boyd, Stephen, and Keromytis, Angelos. "SQLrand: Preventing SQL injection attacks". In Proc. of the 2nd Applied Cryptography and Network Security. Conf. (ACNS 2004), pages 292–302, Jun. 2004.
12. Chaofeng, L., 2006. Research and Development of Data Preprocessing in Web Usage Mining .In the Proceedings of International Conference on Management Science and Engineering , 1311-1315.




*Published By:*
*Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)*