A Proposed Methodology on Predicting Visitor's Behavior based on Web Mining Technique

Abdel Karim Kassem[1]
University of Angers
Angers, France

Bassam Daya[2]
Lebanese University
Saida, Lebanon

Pierre Chauvet[3]
Université Catholique de l'Ouest
Angers, France

***Abstract*—The evolution of the internet in recent decades enlarge the website's reports with the records of user's activities and behaviors that registered in the web server which can be created automatically in the web access log file. The feedback concerning the user's activities, performance and any problem that may be occur including the cyber security approaches of the web server represents the principal raison of applying the web mining technique. In this paper, we proposed a methodology on predicting users behavior based on the web mining technique by creating and executing analysis applications using a Deep Log Analyzer tool that applied on the web server access log of our faculty website. Furthermore, an associated programmed application has been developed which employs the extracted data into dynamic visualizations reports(tables, graphs, charts) in order to help the web system administrator to increase the web site effectiveness, we had creating a suitable access patterns that permits to identify the interacting users behaviors and the interesting usage patterns such as the occurred errors, potential visitors, navigation activities, behavioral analysis, diagnostic study, and security alerts for intrusion prevention. Moreover, the obtained results achieved the aim of producing a dynamic monitoring by extracting investigation summaries which analyses the discovered access patterns that registered in the faculty web server in order to improve the web site usability by tracking the user's behaviors and the browsing activities. Our proposed tool will highlight providing a security alerts against the malicious users by predicting the malicious behaviors taking into consideration all the discovered vulnerabilities by detecting the corrupted links used by the abnormal visitors.**

***Keywords*—Web server; log file; web mining; behavior; pattern; web usage mining; visualizations; vulnerabilities; security**

I. Introduction

Web Usage mining is the strategy of applying web mining techniques to discover and analyze in real time clickstreams usage patterns and related data generated as a result of user interactions with one or multiple web sites. Specifically, web usage mining is the process of grabbing and extracting valuable information in order to find patterns relating to user's behavior of a specific web based system that can determine: who they are, and what they tend to do. Web usage mining techniques consists of the following sections: pre-processing, pattern discovery, pattern analysis.

When a user requested specific and particular resources of web server, each request will be recorded and stored in the web log. This record is referring to the browsing behavior of the user. In Web Usage Mining, data can be collected from multiple resources such as: files (image, sound, video and web files), operational databases and server log files that can include web server access logs and application server logs. Otherwise the collected data in the web log file will be an unstructured format and it can't be used directly for mining purposes, many techniques should be applied on it, the Pre-processing technique play the role of converting the data into suitable and organized form that can helps to precise the pattern discovery and to provide accurate, appropriate and summarized information for data mining intent. Data pre-processing, includes data cleaning, user identification, user sessions identification, path completion and data integration. Pattern discovery benefit from the preprocessing results in order to offer some techniques such as statistical analysis, sequential pattern analysis, association rules, clustering and other techniques. The pattern analysis should be executed and performed by the following techniques: visualization techniques, OLAP techniques and usability analysis.

Aside from detecting the visitors' activities and their behavior, web usage mining can be effectively used to detect existing weaknesses on the web server components and analyzing audit results for anomalous patterns detection.

This research is divided into two parts, the first one by proposing a methodology based on the web usage mining technique that can easily detect the visitor behavior by analyzing the registered visitors' activities on the log file and exporting analysis results to describe the usability of the faculty website, the second one is to discover the cyber-attacks by monitoring the visitors through the links sent to our web server.

To achieve our target, we apply the web usage mining by selecting the data type from our university apache web server which generates the web log file that used for mining purposes, these techniques are used to facilitate the determination of the user behavior and their activities on the web server by creating the rule of the access patterns selection, Furthermore, we will focus on generating some summaries in order to highlight the occurred errors that can be happened on our faculty website, analyzing the traffics, controlling the accessed web server resources, and detecting the illegal activities for expected visitor by controlling the accessed links to discover the web page vulnerabilities which it is a weakness that can be exploited by a threat attacker in order to perform cybercrime actions on the web server.

This paper is organized as follows. Related work in Section 2. Section 3 define the web usage mining methodology and

providing an overviews about its types. Section 4 presents the web usage mining techniques according to the behavioral detection approaches. Section 5 describes our proposed methodology of the detection, followed by Section 6 where we state our experimental results and the extracted analyses summaries. Section 7 conclude the research work which its supported by a proposed perspective that can involve this research topic.

II. Related Works

In this section, we reveal the related work concerning our research study area. The daily web usage of websites with the big amounts of data resulted every second derive us to conclude that much attention has been drawn to the web usage mining that represents one of the popular research areas.

In web usage mining, data analyzation is essential for tracking the user behavior in order to serve the users in efficient way.

Several researchers that are shown in [1][2] developed preprocessing data model; they collected the data related to user ID, path completion, session ID, transaction ID etc. In this way they improve the organization by facilitating the determination of particular clients, products marketing plans and other promotional goals, etc.

According to [3], the authors presents the web log data files and their data difficulties. In addition, the author highlighted about the lessons and metrics based on e-commerce and about the web server's insufficiencies then he introduces some statistical graphs to find the fitting solutions and cover the resulted issues.

The authors in paper [4] presented a technique for detecting the interests of the visitors according to a study of the site-keyword graph. This technique can extract sub-graphs to reveal the major interests of the users taken from the site-keyword-graph were the data is collected from the log data of the website. According to [5] the authors described a mining algorithm for incremental web traversal pattern, this algorithm employs the mining results and predicts another patterns using the deleted or inserted data parts of the logs in the websites like the mining duration that may be reduced. The authors present in [6] an analysis on the web log data via a method for statistical analyzation. Moreover, this author clarifies a recommended tool for efficient realization and interpretation of the preprocessed statistical results taken from log file.

According to [7], the authors worked on this research topic by abstracting the log lines to log event types in order to mine the system logs, this work has been accomplished by presenting a technique based on clustering using the simple log file clustering tool to abstract the logs; moreover, this technique is useful when we cannot access the source code of the application. This research was done by the virtual computing lab at the university of North Carolina state.

These papers [8][9] explore the user session by applying detailed characterization study, after that the authors preview the results for several views such as each user requests per session, page number requested per each session, the session length.

III. Web Usage Mining

Web mining consist of three categories: Web content mining, web structure mining and web usage mining. The concept of Web usage mining is to gather data and information generated by the web. While the concept of the web content and structure mining is to apply the primary data on the web, moreover web usage mining will mine the secondary data obtained by the interactions of the multiple users in the web[10]. One of the functions of the web usage mining is to include the data from the web server access logs, browser logs, proxy server logs, registration data, user profiles and sessions, user queries, cookies, mouse clicks and scrolls, bookmark data and other detailed data as interaction results.

The web usage mining technique can be declared by three steps process: data pre-processing, pattern discovery and pattern analysis as we shown in the Fig. 1.

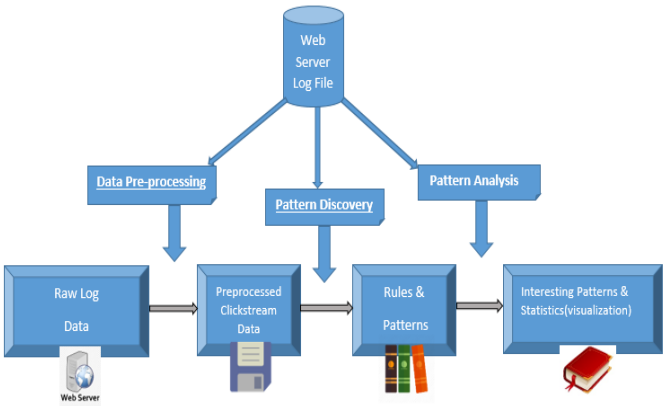

Fig. 1. Web usage Mining Process.

A. Data Preprocessing

By accessing any website, actually the user's behaviors[11] will be stored in the web server log file in unclear and unorganized form. As a definition, data preprocessing is the process for converting the raw data presented in log files into suitable form such as data base or different data store type which contribute effectively when applying the data mining algorithm. Since the main log file cannot be directly used in the web usage mining process, due to the large amount of irrelevant entries in the log file and difficulties and many reasons. Hence, web log file's preprocessing becomes essential and significant. Nowadays, many researches centers are interested in data preprocessing of Web Usage Mining methodology.

Thus, data preprocessing plays an essential role in increasing the mining accuracy in order to improve the data quality for further usage.

B. Pattern Discovery

Pattern discovery employs the preprocessing results to offer some techniques such as statistical analysis, association rules, sequential pattern analysis, dependency modeling,

classification and clustering to capture beneficial useful information. The results that has been grabbed can be represented and employed in several ways such as graphs, charts and tables, etc. for example the visitor's location can be specified using his own IP address. Therefore, by discovering the web visitors[12], the web server administrator can detect the most active countries who's visiting a certain website or any web page that can provide the useful information relevant to the specific country.

C. Pattern Analysis

Pattern analysis can be classified as the final step in the Web Usage Mining process. The main purpose of applying the pattern analysis[13] is to filter out the unusable and the non-beneficial rules and patterns from the set that has been found in the pattern discovery phase. Most Pattern analysis techniques are used to attain the above mentioned purpose.

One of the above techniques is the knowledge query mechanism like SQL which is a standard language for storing, retrieving and manipulating data in databases[14]. Another method is called (OLAP) which is an operation to load usage data into a data cube in order to perform Online Analytical Processing. Visualization techniques is the process of conveying information in a way that the information can be quickly and easily digested by the viewer or the analyzer such as graphing patterns by assigning colors to a specific value in order to highlight overall patterns in the data. Content and structure information are used to extract patterns that contain several pages of a certain usage type that can match with a certain hyperlink structure.

IV. Web usage Mining and Behavioral Detection Approaches

Web mining is an application of data mining methodology that discovers the usable patterns from the internet according to the World Wide Web protocol. As the name inspires, by using the web mining techniques, this information will be gathered from the internet. This technique uses automated devices that reveal and extract data from the web servers and much reports on the internet that permits the companies and educational organizations to extract structured, semi structured and unstructured data from browser actions, server logs, website, web page's contents, page Links and another sources [15]. Web mining techniques[16] can be applied also to detect the user activities as shown in the Fig. 2; this can be reached when we employ their techniques to discover the user behavior as well as it is used to handle the problems presented in the databases and the cyber security troubles through analyzing the illegal and the irregular user activities.

Web usage mining is the practice of extracting valuable information from the server logs in order to find and conclude what visitors are looking for in the interconnected networks(internet), after that the discovered knowledge by the visitors are taken to roam and navigate via the websites [17]. In this paper, we proposed a "mixture approaches" the concept of web usage mining is used intended for the visitor's behavior detection. We can discover the web visitors' information that derive us to identify the user's movements and activities in order to detect and analyze the web traffics, the occurred errors, the users' activities, the abnormal and illegal actions and the security approaches.

The main advantage of the web usage mining technique is to propose a series of those combined approaches that exclusively save the time as well as decrease the estimated cost. Using this kind of techniques, the web administrator will dynamically analyze the user activities and the human efforts to extract the desired reports will be reduced and there will be no need to hard physical potential during the detection.

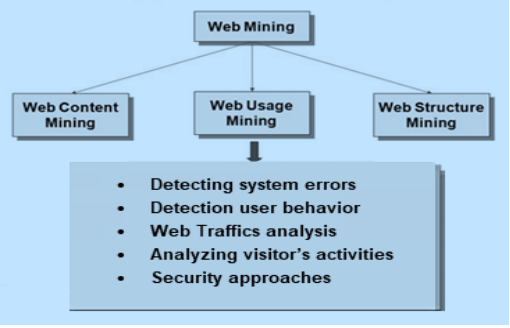

Fig. 2. The Behaviral Detection Approches based on the Web usage Mining Technique.

V. Proposed Methodology

Web Usage mining is the process of applying web mining techniques to discover the approaches of usage patterns from the extracted Web data. The web usage mining is one of the significant and fast developing zone of web mining that it is considered as an important part of the advanced technology (web mining) to discover the user's behaviors events. In this research paper, we developed and applied the Deep Log Analyzer tool associated with a programmed application that requires the web server log file to create a suitable pattern according to the visitor's behaviors by generating statistical and web usage mining reports which can analyze all the detected behaviors approaches.

In this section, we propose the used methodology that assists the web administrator to analyze the occurred system errors, security alerts and user's activities by detecting their behaviors on the web server logs. The steps bellow is included in the proposed methodology.

A. Data Collection

In this section, we present the data collection that applied in our research that has been extracted from our faculty web server access log. The web log stores the visitor's activities per each user visit and hit. The collected data was extracted from log file during a period of four days on February 2018 as shown in the TABLE Ibelow.

TABLE I. Details of the Input Data (Access Log File)

Access Log File Details	
File Name	iut.ul-iut.net-Feb-2018
Period	23 Feb 2018 – 26 Feb 2018
Size(KB) after preprocessing	1852.8
Number of entries	6742

B. Data Selection

In this section, we present the data selection concept that we used. Absolutely the web mining methodology has three kinds of data: the server side data, the middle data (proxy side) and the client side data. In our work, we employed the case of web server use.

C. Web Server Log

A web server refers to computer or to server software or both of them working together to transfer web pages. The web server uses HTTP (Hypertext Transfer Protocol) in order to serve the web server files that form Web pages to web users directly in response to achieve their requests, which are forwarded by their HTTP clients the main log file cannot be directly used in the web usage mining process. Log files[18] are files which are composed, established and maintained in a web server. Every hit to the Web site by the users, including each view of HTML documents, images or any other object will be logged. The raw web log file format is ultimately formed of single line text for each interaction, mainly it is a hit related to the web page interactions. The log files have the capability to maintain different types of information [19] and it will be presented in the log file and should summarized who, where and when the users visited the website [19], and it will serve to discover their behaviors and movements. Moreover, when the users communicate and interact with any website, the interaction's details and the request activity resulted by the web visitor events will be automatically recorded and stored in the web server log file [20][21].

The basic information recorded and discovered in the log file can be shown as

- Username: This identifier will discover who visits the website. The identification of the user principally would be the IP address.
- Visiting Path: The path that the user typed while visiting the website.
- Path Traversed: it will distinguish the path taken by the user via different links.
- Time stamp: The time duration when the user spends on each web page while surfing through the website, this record recognized as a session.
- Last visited Page: The visited web page by the users before the leaving.
- Success rate: The number of downloads made and the number of replicating activities experienced by the user that can specifies the success rate of the website.
- User Agent: This is the browser that can indicate from where the user sends the request to the web server. It will be formed as a string that characterizes the type and the version of browser software being used.
- URL: It will be the resource of the user access. It may be an HTML page, a CGI program, or a script.
- Request Type: The method chosen for transferring data such as GET, POST

D. Tool Selection

Most of the valuable information about any website visitor stored in the log file on the web server, after analyzing these data we can generate beneficial reports as summaries, graphs and analytical figures by using the web usage mining technique which it can be done using various tools. A variety of tools are available in the internet assists the web administrator to apply analysis tasks by accessing the web server log files which produces effective web usage mining reports as output. Some of the most widely used tools are: Google analytics, webalizer, W3Perl, and AWStats. In this paper, we select our deep log analyzer with an analytical application to analyze the desired goal by examining the log file in order to achieve the target, this can facilitate of obtaining an output as reports about the accessed information, user behavior analysis, system errors, threatened links, security approaches, user identity, time, zone, URL, browser and OS of the users. Unlike other tools, our tool has the ability to analyze different types of logs including FTP logs. It can analyze the web site visitors' behavior to get the complete usage statistics that improve the usability and stability of our web site and provide an analytical protective studies in order to avoid the web vulnerabilities that actually occur on the web server.

We can study the extracted results and generate the following reports according to its own features

- View reports about accessed site's resources
- View reports about the most visited links
- View investigation report of the user behavioral activity
- Monitor and control the illegal visitors'
- View the abnormal sites that refers to the web traffic
- Reveal the search queries and search engine spiders
- Reveal the user browsers and operating systems
- View the web server errors
- Apply comparative reports in different time periods
- Analyze the log file with respect to all popular web servers such as IIS on Windows or Apache on Unix/Linux

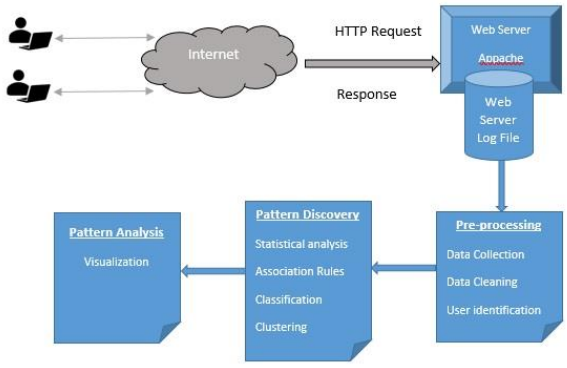

Fig. 3. The Proposed Methodology.

E. Methodology Implementation

Usually, by clicking on a web link or any click stream by the visitors, the web server stores and generates these actions in the log file. Log file consists of multiples raw records about all web pages that provide the discovery detection[22] of the user's behaviors. This paper sheds the light on pattern analysis of the visitor taken from the log data of our university web server.

Throughout this research paper, we can illustrate our framework in the Fig. 3 by analyzing the proposed methodology that permits to understands and evaluate the web visitor's behavior. Hence the user uses the internet service to serve web pages until he/she reach our faculty's website whether directly, or by using the search engines or through referrals resources. The user's actions on the website will be stored on apache server log file.

By applying the Web usage mining, we can collect and investigate the recovered data from it. Furthermore, the next step is to deal with user's interaction through the website in order to infer their behavioral patterns and profiles.

The main purpose of our research is to detect the information with respect to the visitor's behaviors. The extracted Information from the log file will be employed in our tool in order to extract web usage mining statistical results

The most important results will be displayed as listed below:

- Links and Resources Analysis
- Server Content Analysis
- Brower Analysis
- Web Page Analysis
- Security Approaches
- Operating System Analysis
- Time and Place Analysis

VI. Experimental Results and Analysis

The main objective of the web usage mining technique is to generate statistical reports as output results that can be used to detect some valuable information after analyzing them, in this paper we focused on the data extraction from our faculty's web server log file as an input concerning the visitors and the user's behaviors in order to generate an investigation reports with respect to the web server status.

Our research will display and discuss several experimental results as:

1) General activity: the main general activities of our faculty web sites are shown below in the fig.4. which clarifies a brief summary about visitor's information during selected dates.

a) Selection information summary during selected dates

The Fig. 4 illustrates a summary report that will be explained below concerning the statistical results with respect to the number of hits, visits, visitors and page views of the faculty's web site.

- Hits summary that includes the number of hits, the number of successful hits and the outgoing and incoming traffic (as total or per day).
- Visits summary that includes the total number of user visits, the average number of visits per day and the average visit duration.
- Visitors' summary that includes the number of unique visitors, the visitors who visited once, the repeated visitors, the average visits per visitor and the most visitors from this country.
- Page views summary that includes the total page views, the most popular page, the most popular downloaded file, the most popular entry page and the most popular exit page.

Hits Summary [Details...]	Total	Per Day
Number of Hits:	318	80
Number of Successful Hits:	266 (84%)	67
Outgoing Traffic:	6.67 Mb	1.69 Mb
Incoming Traffic:	204 Kb	51 Kb

Visits Summary	Total
Number of Visits :	51
Average Number of Visits per Day:	13
Average Visit Duration :	7:44 Min

Visitors Summary	Total
Number of Unique visitors :	44
Visitors who visited once:	39 (89%)
Repeat visitors:	5 (11%)
Average Visits per visitor :	1.16
Most visitors from this Country :	Lebanon (36% visitors)

Page Views Summary		Hits
Total Page Views :		204
Most popular Page :	.../wp-login.php??goto=99999...	15
Most popular Download :	.../33adbf46c4fd78664e915ffa...	9
Most popular Entry Page :	/	6
Most popular Exit Page :	/	5

Fig. 4. The Proposed Methodology:A Brief Summary about the Visitor's Information.

b) Referral summary information

The Fig. 5 as we shown below represents and concludes the referral and search engine summaries.

Referral Summary		Hits
Top Referring Website :	http://www.iut.ul.edu.lb	103

Search Engines Summary		Hits
Top Search Engine :	Google	55
Top Key Phrase :	NOT PROVIDED	55
Spiders Requests :		1

Fig. 5. The Referral Information about the Website.

- Referred summary includes top referring websites on the web server.
- Search engine summary includes top search engine that provide the users to access the university website, top key phrase and spider requests on the search engine provider.

c) Technical information

The Fig. 6 reveals a technical summary that contains the most popular browser, the most popular operating system and the error hits that happened on the web server.

Technical Summary	
Most Popular Browser :	Mozilla or other Mozilla based 5.0
Most Popular Operating System :	Android
Error Hits :	52 (16%)

Fig. 6. Technical Summary for the Website.

2) Visitors activities: by controlling the visitor's activities on the web server, many difficulties can be encountered to detect the visitor's behavior and their purposes. After employing the web usage mining techniques. We achieve the target and we will be able to detect valuable information about the top visitors with their countries and the number of visits that contacts the concerned website as well as the daily and hourly user activities facts that occurred on the web server log file.

a) Selection information summary during selected dates:

Visitor	Country	Number of Visits
78.40.182.55	Lebanon	4
92.241.42.91	Jordan	2
41.227.59.60	Tunisia	2
41.223.201.249	Sudan	2
146.185.35.144	Lebanon	2
46.229.168.72	Netherlands	1
46.229.168.71	Netherlands	1
46.229.168.70	Netherlands	1
46.229.168.68	Netherlands	1
46.229.168.67	Netherlands	1
46.229.168.65	Netherlands	1
42.61.41.114	Singapore	1

Fig. 7. Top Detected Visitors Of The Faculty Website.

The Fig. 7 represents the most active visitors identified by their IP addresses, the countries and visit's numbers of the website.

b) Visitors spending Time

The graph bellow represented by the Fig. 8 determines the spending period time of the visitors in our faculty website. The x-axis represents the spending average time of the visits. however, the y-axis indicates the total number visits of each visitor. We can conclude from this statistical graph that the spending time is continued as long as the web server receives hits from that visitor.

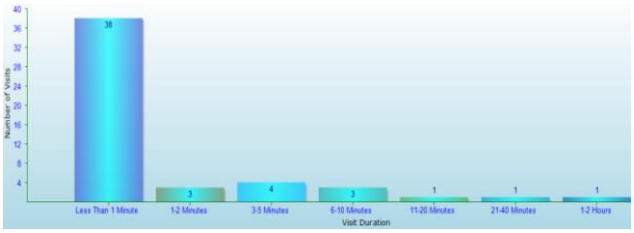

Fig. 8. Visitor's Spending Time on the Faculty Website.

Day of Week	Number of Hits	Data Transferred (Kb)
Sunday	25	83
Monday	122	1,039
Tuesday	155	5,886
Saturday	16	0

Fig. 9. Popular Days of the Week with the Number of Hits and Data Traffic.

c) Visitors daily activity

The Fig. 9 gives a clear image how the traffic may vary from a day to another in the same week. The traffic is presented by the hits number of each visitor measured in Kb as the transferred data related to the users, this figure will reveal about the days that the website achieves the traffic as a quantitative indicator about the exchanging data in the web server.

d) Hourly rate activity

The Fig. 10 below displays the traffic on the website that can be changed depending on the daily traffic time, we can find out the hourly time of a day of each hit on the website measured in Kb to display the transferred data related to the users.

Hour of Day	Number of Hits	Data Transferred (Kb)
4	28	189
5	8	42
6	36	520
7	54	316
8	31	867
10	36	4,496
11	73	185
12	48	366
16	4	28

Fig. 10. Popular Hours of the Day Ranked by the Transferred Data.

e) Number of the visits by the visitor

The Fig. 11 shows the number of visits for each visitor in order to concluded the visitors' loyalty and interest according to the number of visitors.

Number of Visits by Visitor	Number of Visitors
1	39
2	4
3-5 Visites	1

Fig. 11. The Number of the Visits Per Visitor.

f) Top visitors countries

The Fig. 12 illustrates the origin countries of the university web visitors ranked by the number of unique visitor from that country.

Country	Number of Unique Visitors
Lebanon	16
Netherlands	13
United States	6
United Kingdom (Great Britain)	1
Tunisia	1
Syrian Arab Republic	1
Sudan	1
Singapore	1
Jordan	1
Italy	1
France	1
Total	43

Fig. 12. Top Visitor's Countries.

g) Browsers

The Fig. 13 displays the web browsers types employed by the visitors ranked by number of hits for each browser that identify the most used ones while accessing our web faculty. Furth more, the data Transferred column in the figure below shows the transferred amount traffic in KB's from each web browser.

Browser	Number of Hits	Data Transferred (Kb)
Google Chrome	155	1,758
Mozilla or other Mozilla based	89	548
Firefox	40	3
Safari	12	84
MS Internet Explorer	12	197
Android	8	4,391

Fig. 13. The Most used Browsers.

h) Operating system

The report bellow represented by the Fig. 14 illustrates the most used browser with the operating system platform which used to access the web faculty. The installed operating system platforms on the visitor's computer should be ranked by the number of hits from each OS. On another hand, the data transferred column shows the traffic amount in KB's transferred to the visitors.

Platform	Number of Hits
Android	119

Browser	Number of Hits	Data Transferred (Kb)
Google Chrome	111	1,454
Android	8	4,391
Total	119	5,845

Platform	Number of Hits
Windows 7	85

Browser	Number of Hits	Data Transferred (Kb)
Google Chrome	44	304
Firefox	25	0
MS Internet Explorer	12	197
Mozilla or other Mozilla based	4	91
Total	85	592

Platform	Number of Hits
Apple iPhone/iPod	57

Browser	Number of Hits	Data Transferred (Kb)
Mozilla or other Mozilla based	51	326
Safari	6	32
Total	57	358

Platform	Number of Hits
Windows 10	6

Browser	Number of Hits	Data Transferred (Kb)
Firefox	6	0
Total	6	0

Platform	Number of Hits
Windows 8.1	4

Browser	Number of Hits	Data Transferred (Kb)
Firefox	4	3
Total	4	3

Platform	Number of Hits
Mac OS X	1

Browser	Number of Hits	Data Transferred (Kb)
Safari	1	52
Total	1	52

Total	272

Fig. 14. The used Operating Systems Accessed by the Visitor's Web Browsers.

1) Access resources control and security approaches:

a) Top downloaded files

The 0 shows the popularity of the downloaded files from the faculty website. Downloads are ranked by the number of files that requested by the visitors (number of hits). This figure shows the downloaded files with their specific extensions. For example, these extensions include zip, exe, rar and tar, etc for compressed file, graphics (gif, jpg, etc.), sound (wav, mp3 ...) and video (avi, mpg, mp4...) otherwise the files that are not considered as downloaded file will not appear in this report.

FileName	Number of Hits	Data Transferred (Kb)
/subCat/33adbf46c4fd78664e915ffa2045c1b5.pdf	9	4,243
/schedules/5d08a3612465501c05e787e07f41501f.pdf	4	269
Total	13	4,512

Fig. 15. Top Downloaded Files.

Directory	Number of Hits	Data Transferred (Kb)
/	180	738
/css/	27	101
/ccne/	19	213
/img/	17	138
/subCat/	9	4,243
/img/news/	7	590
/css/fonts/	7	0
/cap/Update/	6	0
/img/events/	5	394
/schedules/	4	269
/js/	4	130
/img/slider/	4	5
/ccne/plugins/iCheck/square/	3	3
/ccne/plugins/bootstrap-wysihtml5/	3	166
/ccne/plugins/iCheck/	3	1
/ccne/plugins/iCheck/flat/	3	3
/ccne/plugins/iCheck/futurico/	3	1
/ccne/plugins/iCheck/line/	3	5
/ccne/plugins/iCheck/polaris/	3	1
/css/themes/	3	3
/ccne/plugins/iCheck/minimal/	3	3
/ccne/c:/	2	0
Total	318	7,007

Fig. 16. The Top Accessed Directories.

b) Accessed directories

The popularity of the web server directories is declared as shown in the figure below [Fig. 16]. This report is ranked by the number of visitors that requested the web pages or any file located in that directory.

The Data Transferred column shows the total number of Kb's transferred by the visitors of the web server according to the visited directories.

c) Search engines

When a user executes an online search query, the search engine will explore via its searchable index and will returns the results that are related to the desired searcher's query. The outputs are ranked based on the popularity of the website that provides the information. The value and the importance of a website is specified by several factors such as the keywords appearance on the web page, the relevancy of the web page content, the quality of hyperlink, the related social elements (such as Facebook, Instagram, Tweeter likes or shares), and other factors. Therefore, the value of studying the requested search engine is to know the access methods to a website that it is very influential in discovering the effective factors in the website search engine optimization. The figure below [Fig. 17] shows a list of search engines requested by the visitors to find the faculty web site ranked by the number of referrals (Number of Hits column) for each search engine.

Name	Number of Hits
Google	55
Bing	42

Fig. 17. Top used Search Engines.

Referral Site	Number of Hits
http://www.iut.ul.edu.lb	103
http://iut.ul.edu.lb	50
https://www.google.com.lb	48
http://m.facebook.com	17
https://www.google.com	5
https://www.google.jo	2

Fig. 18. Referring WebSites.

d) Referrals website

The Fig. 18 displays the referrer websites that may help to drive the external visitors to our website. These websites ranked by the number of hits received from that referrer.

e) Security alert

Providing a website security, mostly controlling the user behavior has become one of the most important concerns of the technological research centers over the past few years. Many academic companies are joining the game in hopes of capitalizing from the research centers to have a secured web server by controlling the accessed resources in it. One of the essential vectors to provide a fundamental security is the Access Resources Control. When we talk about the access control, the researchers must be concerned with respect to the mechanisms to restrict access to a resource. We have to take into consideration about who are the visitors that connect to our website in order to detect the visitors behavioral by controlling the viewed and visited pages as well as all the accessed resources. The figures as shown below 0 and Fig. 20 will detect the popularity of the viewed and visited web pages that ranked by the number of hits and the transferred data of requested pages by the visitors that will highlight the importance of controlling the URL type and the structure form in order to detect the irregular resources (url, page, directory)accessed by the user according to the main resources, as well as extracting a summary about the quality of the visited resources concerning the web pages in order to classify the fearing of that visitors were its behavior can be detected and determined by studying the irregular URL cases (Sql injection,

XSS, SSRF, Directory Traversal) ranked by the detected behavior type.

a) Diagnostic

Mainly, the practices of the web usage mining techniques play an essential methodology in tracking the visitor's activities and its relation with respect to the other networks. Web system administrator employs this kind of techniques in the log file in order to monitor the desired network and the web server errors that can permits the identifying of the vulnerabilities that may happen in the web server to access critical and important information known as the cyber security attacks. Moreover, our proposed tool plays the role of detecting the occurred errors using the regular expression technique After analyzing the presented errors, we are able to identify who can play the illegal activities on the web server.

We can conclude from the figure below [Fig. 21] that the error "404" is the most error that occurred on the web server; moreover, we can observe the targeted pages in order to determine and find the best solution to fix the discovered vulnerabilities.

FileName	Number of Hits	Data Transferred (Kb)
/wp-login.php??goto=999999.9+%2f**%2fuNiOn%2f**%2faLl+%2f**%2fsElEcT(%2f**%2fsElEcT+%2f**%2fcOnCaT(0x217e21,count(t.%2f**%2ftAbLe_nAmE),0x217e21)+%2f**%2ffRoM+information_schema.%2f**%2fsChEmAtA+as+d+join+information_schema.%2f**%2ftAbLeS+as+t+on+t.%2f**%	15	0
/ccne/adminPFE.php	11	203
/??goto=SELECT%20CHAR(0x66)	11	76
/??goto=10%3B%20DROP%20TABLE%20members%20%2F*	11	76
/??goto=ASCII()	11	76
/??goto=10%3B%20DROP%20TABLE%20members%20--	11	76
/	10	69
/schedule.php	10	35
/subCat/33adbf46c4fd78664e915ffa2045c1b5.pdf	9	4,243
/favicon.ico	8	0
/cap/Update/update_Schedule.php	6	0
/css/ajax-loader.gif	6	0
/img/logouni.png	6	113
/img/course-nav-prev.png	5	8
/img/course-nav-next.png	5	9
/img/news/007f4422189933377c4cc893adfc4b79.jpeg	5	383
/wp-login.php??goto=999999.9+%2f**%2fuNiOn%2f**%2faLl+%2f**%2f%2f**%2f**%2%2f**%2sElEcT(%2f**%2fsElEcT+%2f**%2fcOnCaT(0x217e21,count(t.%2f**%2ftAbLe_nAmE),0x217e21)+%2f**%2ffRoM+information_schema.%2f**%2fsChEmAtA+as+d+join+information_schema.%2f**%2ftA	5	0
/wp-login.php??goto=999999.9+%2f**%2%2f**%2%2f**%2fuNiOn%2f**%2faLl+%2f**%2fsElEcT(%2f**%2fsElEcT+%2f**%2fcOnCaT(0x217e21,count(t.%2f**%2ftAbLe_nAmE),0x217e21)+%2f**%2ffRoM+information_schema.%2f**%2fsChEmAtA+as+d+join+information_schema.%2f**%2ftAbLeS+as	5	0
/schedules/5d08a3612465501c05e787e07f41501f.pdf	4	269
/css/fonts/slick.woff	4	0
/style.css	4	27
/schedule.php?file=/etc/	4	14
/schedule.php?javascript%3Aalert%28%27%27%29	4	14
/schedule.php?javascrip:alert('')	4	14
/ccne/adminPFE.php?Login='%20and%20'1'='1~~~&Password='%20and%20'1'='1~~~&ret_page='%20or%20'1'='1~~~&querystring='%20%2B%20(SELECT%20FieldName%20FROM%20TableName%20LIMIT%201,1)%20%2B%20'~~~&FormAction=login&FormName=Login	4	5
/img/slider/navigation-icon.png	4	5
/index.php?Login='%20and%20'1'='1~~~&Password='%20and%20'1'='1~~~&ret_page='%20and%20'1'='1~~~&querystring=';%20EXEC%20master..sp_makewebtask%20""\\10.10.1.3\share\output.html"",%20';%20SELECT%20*%20FROM%20INFORMATION_SCHEMA.TABLES""~~~&FormAction=login&F	4	28
/js/jquery.min.js	4	130
/schedule.php??file=%3Cscript%3Ealert%28%27%27%29%3C%2Fscript%3E	4	14
/schedule.php?file=/etc/shadow	4	14
/ccne.php??goto=AND%20%20%27a%27%3D%27a%27	4	28
/css/themes/default-theme.css	3	3
/css/slick.css	3	4
/ccne/plugins/iCheck/square/_all.css	3	3
/ccne/plugins/iCheck/polaris/polaris.css	3	1
/ccne/plugins/iCheck/minimal/_all.css	3	3
/ime.php	3	16
/css/jquery.tosrus.all.css	3	7
/ccne/plugins/iCheck/futurico/futurico.css	3	1
/ccne/plugins/iCheck/flat/_all.css	3	3
Others	89	1,036
Total	318	7,006

Fig. 19. Top Accessed Pages and Resources.

Access Pattern	Behaviors Detection
/wp-login.php??goto=999999.9+%funNiOn%f**%2fal+%2f**%2fsElEcT(%2f**%2fsElEcT+%2f**%2fcOnCaT(0x217e21,count(t.%2f**%2ftAbLe_nAmE),0x217e21)+%2f**%2ffRoM+information_schema.%2f**%2fsChEmAtA+as+d+join+information_schema.%2f**%2ftAbLeS+as+t+on+t.%2f**%	Malicious Authentication
/??goto=10%3B%20DROP%20TABLE%20members%20%2F*	Database Threat
/schedule.php?javascript%3Alert%28%27%27%29	Client Side Threat
/ccne/adminPFE.php?Login='%20and%20'1'='1~~~&Password='%20and%20'1'='1~~~&ret_page='%20or%20'1'='1~~~&querystring='%20%2B%20(SELECT%20FieldName%20FROM%20TableName%20LIMIT%201.1)%20%2B%20'~~~&FormAction=login&FormName=Login	Malicious Authentication
/index.php?Login='%20and%20'1'='1~~~&Password='%20and%20'1'='1~~~&ret_page='%20and%20'1'='1~~~&querystring=';%20EXEC%20master..sp_makewebtask%20"\\10.10.1.3\share\output.html""".%20';%20SELECT%20*%20FROM%20INFORMATION_SCHEMA.TABLES""~~~&FormAction=login&20SELECT%20*%20FROM%20INFORMATION_SCHEMA>TABLES"""~~~&FormAction=login&F	Malicious Authentication
/ccne.php??goto=AND%20%20%27a%27%3D%27a%27	Malicious Query String

Fig. 20. Some of the Detected Behaviours about the Malicious Visiotrs.

Error	Number of Hits
404 - File Not Found	52

Page	Number of Errors
/wp-login.php??goto=999999.9+%2f**%2fuNiOn%2f**%2faLl+%2f**%2fsElEcT(%2f**%2fsElEcT+%2f**%2fcOnCaT(0x217e21,count(t.%2f**%2ftAbLe_nAmE),0x217e21)+%2f**%2ffRoM+information_schema.%2f**%2fsChEmAtA+as+d+join+information_schema.%2f**%2ftAbLeS+as+t+on+t.%2f**%	15
/favicon.ico	8
/css/ajax-loader.gif	6
/wp-login.php??goto=999999.9+%2f**%2fuNiOn%2f**%2faLl+%2f**%2f?%2f**%2%2f**%2%2f**%2sElEcT(%2f**%2fsElEcT+%2f**%2fcOnCaT(0x217e21,count(t.%2f**%2ftAbLe_nAmE),0x217e21)+%2f**%2ffRoM+information_schema.%2f**%2fsChEmAtA+as+d+join+information_schema.%2f**%2ftA	5
/wp-login.php??goto=999999.9+%2f**%2%2f**%2%2f**%2fuNiOn%2f**%2faLl+%2f**%2fsElEcT(%2f**%2fsElEcT+%2f**%2fcOnCaT(0x217e21,count(t.%2f**%2ftAbLe_nAmE),0x217e21)+%2f**%2ffRoM+information_schema.%2f**%2fsChEmAtA+as+d+join+information_schema.%2f**%2ftAbLeS+as	5
/css/fonts/slick.woff	4
/css/fonts/slick.ttf	3
/ccne/c:/arbre%20de%20Haffman	1
/ccne/c:/Abre%20de%20Huffman	1
/apple-touch-icon-precomposed.png	1
/apple-touch-icon-120x120-precomposed.png	1
/apple-touch-icon-120x120.png	1
/apple-touch-icon.png	1
Total	52
Total	52

Fig. 21. The Occurred Web Server Errors.

VII. Conclusion

Nowadays, the Website is considered as the most used means by the internet visitors to collect desired and valuable information. Therefore, the usability and security of a web server resources are very important to provide a website more popular among its visitors.

In this paper, we proposed a methodology on predicting user behavior based on the web mining technique. Our target has accomplished by developing and applying our suggested tools that merge two separate techniques: the web usage mining technique and the cyber security approaches. Through these mechanisms, we can have a potential to create a suitable access patterns in order to detect and identify the system errors which occurred on the web server, as well as the identification of the user's behaviors and their important activities, the potential visitors, the navigation activities and a diagnostic study. Moreover, we can predict the pages that make the errors which can help us to detect the vulnerabilities on our web server, this can be done by controlling all the URL's, directories and web pages in order to provide a security mechanism with respect to the malicious users, specially the abnormal activities, by taking into consideration all the breaches which can be occurred on the faculty web server. Furthermore, our extracted results will be shown as summaries, tables, figures and charts that can be consider as a guide report which discuses each discovered pattern and its behavior, thus can be helpful for the System Administrators, Web Analysts and Website Maintainers to improve and enhance the usability and the security stability of the web server concerning their resources.

As perspective works, our aim is to divide the future work into two parts. The first part will enhance the detection of the visitor behavior, this can make our research more effective by tracking only the actions of interest activities which cause the errors, thus we will discover the accessed pattern in less period of time and with minimum memory system utilization.

On other hand, the second part will focus deeply on the cyber security approaches, it will be released by extending the data extraction period time in order to cover big amount of vulnerable data, this can help to provide a data set in order to develop an intelligent model using the machine learning algorithms that predict the abnormal visitors and the expected attacks.

References

[1] Zhang Huiying, Laing Wei "An Intelligent Algorithm of Data Pre-processing in Web Usage Mining" Proceedings of the 5th world Congress on Intelligent Control and Automation, June15-19, 2004 Hangzhou, P.R.China.

[2] Doru Tanasa et.al Advanced data preprocessing for inter sites Web Usage mining IEEEE computer society 2004.

[3] Kohavi "Mining E-Commerce Data: The Good, the Bad, and the Ugly" KDD 2001, Aug 26-29, San Francisco, CA. Copyright 2001 AC.

[4] Tsuyoshi Murata and Kota Saito "Extracting Users Interests from Web Log Data" Proceedings of the 2006 IEEE/WIC/ACM International Conference of Web Intelligence (WI 2006 Main Conference Proceedings) (WI'06) 2006 IEEE.

[5] Show-Jane Yen, Yue-Shi Lee and Min-Chi Hsieh. "An Efficient Incremental Algorithm for Mining Web Traversal Patterns" Proceedings of the 2005 IEEE International Conference on e-Business Engineering (ICEBE'05).

[6] A-Nikos Koutsoupias "Exploring Web Access Logs with Correspondence Analysis" 2nd Hellenic Conf. on AI, SETN-2002, 11-12 April 2002, Thessaloniki, Greece, Proceedings, Companion Volume, pp. 229-23.

[7] Meiyappan Nagappan, Malden A.vouk “Abstracting log lines to log event types for mining software system logs” 7th IEEE Working Conference on Mining Software Repositories (MSR 2010),2010.

[8] Bettina Berendt, Bamshad Mobasher, Myra Spiliopoulou and Jim Wiltshire “Measuring the Accuracy of Sessionizers for Web Usage Analysis” KDD'99 Workshop on Web Usage Analysis and User Pro_ling WEBKDD'99, San Diego, CA, Aug. 1999. ACM. Springer, LNCS series.

[9] Maristella Agosti Giorgio Maria Di Nunzio “Web Log Mining: A Study of User Sessions “UNIVERSITY OF PADUA Department of Information Engineering.10th DELOS Thematic Workshop on Personalized Access, Prole Management, and Context Awareness in Digital Libraries Corfu, Greece, 29-30 June 2007.

[10] Tsuyoshi Murata and Kota Saito “A Survey on Predicting User Behavior Based on Web Server Log Files in a Web Usage Mining” International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), 2016.

[11] A. Deepa, P. Raajan, "An efficient preprocessing methodology of log file for Web usage mining", NCRIIAMI - National Conference on Research Issues in Image Analysis and Mining Intelligence, 2015.

[12] Ankita Kusmakar, Sadhna Mishra Web Usage Mining: A Survey on Pattern Extraction from Web Logs” International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 9, September 2013 ISSN: 2277 128X,Page:-834-838.

[13] Nina, Shahnaz Parvin, Mahmudur Rahman, Khairul Islam Bhuiyan, and Khandakar Entenam Unayes Ahmed. "Pattern discovery of web usage mining." In Computer Technology and Development, 2009. ICCTD'09. International Conference on, vol. 1, pp. 499-503. IEEE, 2009.

[14] Dhawan, Sanjeev, and Swati Goel, "Web Usage Mining: Finding Usage Patterns from Web Logs." American International Journal of Research in Science, Technology, Engineering & Mathematics (2013): 203-207.

[15] A. Saluja, B. Gour, L. Singh, "Web Usage Mining Approaches for User's Request Prediction: A Survey", IJCSIT-International Journal of Computer Science and Information Technologies, vol. 6, no. 3, 2015.

[16] Sudha Nagesh, “Roll of Data Mining in Cyber Security”, Journal of Exclusive Management Science –May 2013-Vol 2 Issue 5 - ISSN 2277 – 5684.

[17] L.K. Joshila Grace, 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.

[18] G. K. Lekeas, “Data mining the web: the case of City University’s Log Files,” 2000.

[19] K. Suneetha and R. Krishnamoorthi, “Identifying user behavior by analyzing web server access log file,” IJCSNS International Journal of Computer Science and Network Security, vol. 9, no. 4, pp. 327–332, 2009.

[20] K. Etminani, “Web Usage Mining: users’ navigational patterns extraction from web logs using Ant-based Clustering Method,” IFSAEUSFLAT, pp. 396–401, 2009.

[21] Ratnesh Kumar Jain, Dr. R. S. Kasana1, Dr. Suresh Jain, “Efficient Web Log Mining using Doubly Linked Tree,” International Journal of Computer Science and Information Security, IJCSIS, vol. 3,2009.

[22] S. G. Langhnoja, M. P. Barot, and D. B. Mehta, “Web Usage Mining Using Association Rule Mining on Clustered Data for,” International Journal of Data Mining Techniques and Applications, vol. 02, no. 01, pp. 141–150, 2013.