A Comparative Study between Web Mining Tools over some WUM Algorithms to Analyze Web Access Logs

Arvind K. Sharma

Abstract— This paper has attempted to provide an up-to-date survey of the rapidly growing area of Web usage mining, which is the demand of current technology. In this paper, we present an overview of the various research areas in Web data mining and then focus on the Web usage mining tools and techniques. Web mining continues to remain as a potential research area in the present scenario. In this context, the various Web usage mining algorithms are discussed and their relative comparison of merits and demerits are also presented and the most appropriate ones are selected based on the characteristics of the data available from the Web server log files. Finally, we have investigated three powerful Web usage mining tools. The use of these tools is also illustrated through the analysis of one case study. The results of Web usage mining need to be visualised in order to assist with their analysis and interpretation.

Index Terms— Web Data Mining, Web logs, WUM Tools

I. INTRODUCTION

TODAY, the Internet is most emerging technology in the world. The terms Internet and World Wide Web are often used in everyday speech without much distinction. The World Wide Web is also known as 'Information Superhighway'. It is a system of interlinked hypertext documents accessed via Internet. However, the Internet and the World Wide Web (WWW) are not one and the same. The Internet is a global system of interconnected computer networks. In contrast, the Web is one of the services that run on the Internet. It is a collection of text documents and other resources, linked by hyperlinks and URLs, usually accessed by Web browsers from Web servers. In short, the Web can be thought of as an application 'running' on the Internet [1]. The use of internet needs to follow some specific protocol that is given by our service provider. The Web is the universal information space that can be accessed by companies, governments, universities, teachers, students, customers, businessmen and some users. In this universal space trading and advertising activities are held. Further, no one actually knows the size of the World Wide Web (WWW), it is reported to be growing at approximately a 50% increase per year. As of early 1998, over 500,000 computers around the world provided information on the World Wide Web in an estimated 100 million web pages. By 1994, there were approximately 500 Web sites, and by the start of 1995, nearly 10,000. By the turn of the century, there were more than 30 million registered domain names. A decade later, more than a hundred million new domains were added. In 2010, Google claimed it found a trillion unique addresses (URLs) on the Web. A website is a lot of interconnected web pages containing images, videos or other digital assets, which are developed and maintained by a person or an organization. Every website is hosted by at least one web server. A web server is a program that, using the client/server model and the World Wide Web's Hypertext Transfer Protocol (HTTP), serves the files that form web pages to web users [2]. The primary function of a web server is to deliver web pages on the request to the clients. It means delivery of HTML documents and any additional content that may be included by a document, such as images, style sheets and scripts. Every computer on the Internet must have a web server program. Two leading web servers are: Apache and Microsoft's Internet Information Server (IIS). The Apache is the most widely used Web server in this technology. Moreover any web user surfs that website user's some information is stored in Web log which resides in the Web server. Web log stores information of the user activity which performed on the website. Web log contains information about User Name, IP Address, Time Stamp, Access Request, Success Rate. Web mining studies, analyzes and reveals useful information from the Web [3]. Web mining deals with the data related to the Web, they may be the data actually present in Web pages or the data concerning the Web activities. Web mining is an area that lately has gained a lot of interested. This is due essentially to the exponential growth of the World Wide Web and its anarchic architecture and also due to the increase of its importance over the people's life. The researchers want to extract information from it, in order to better understand and to improve its features. The rest of paper is organized as follows: Section 2 describes research phases and applications of web usage mining. Section 3 gives an overview on web logs with their types and location. Section 4 summaries about Literature Survey. Section 5 specifies the methodology. Section 6 reports our experimental results. In section 7 we conclude this paper with summary and describe an outlook for future work. Finally, in the last section references are mentioned.

II. WEB DATA MINING

A. An Overview

Web mining is a very hot research topic that combines two of the activated research areas: Data Mining and World Wide Web. Web mining is a natural combination of data mining and the World Wide Web (WWW).

Manuscript received on May 28, 2012.

Arvind K. Sharma, **Ph.D Computer Science Research Scholar**, School of Engineering & Technology, Jaipur National University, Jaipur, INDIA, (HOD Computer Science, DAV Kota),

It may be defined as the discovery and analysis of useful information from the World Wide Web (Cooley, 1997). As many researchers believe, that the term of Web Mining is firstly proposed by a researcher Oren Etzioni in his paper [4] in 1996. In this paper, he proved the Web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and Web services. Many of the following researchers cited this explanation in their research works. Now days, with the tremendous growth of the data sources available on the World Wide Web and the dramatic popularity of e-commerce in the business community, Web mining has become the focus of quite a few research works. Some of the commercial consideration has presented on the schedule. In both [4] and [5], they suggested a similar way to decompose Web mining into the following subtasks:

(i) **Resource Discovery:** the task of retrieving the intended information from the Web.

(ii) **Information Extraction:** automatically selecting and pre-processing specific information from the retrieved Web resources.

(iii) **Generalization:** automatically discovers general Web patterns at both individual Web sites and across multiple Websites.

(iv) **Analysis:** analyzing the mined Web patterns.

In brief, Web mining is a technique to discover and analyze the useful information from the Web data. The authors of [6] have stated the Web involves three types of data: data on the Web (content), Web log data (usage) and Web structure data. The authors of [7] have categorized the data type: as Content data, Structure data, Usage data, and User profile data.

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents/services(Etzioni,1996). Web usage mining is categorized into three categories[11]:

(i) Content mining (Examines the content of Web pages as well as results of Web searching)

(ii) Structure mining (Exploiting hyperlink structure)

(iii) Usage mining (Analyzing user Web navigation)

Web mining is a term applying data mining techniques to web access logs [8]. Data mining is a non-trivial process of extracting previously unknown and potentially useful knowledge from large databases [9]. It is a technique that can be used to analyse large amounts of data in order to discover patterns. We know that Web data is unstructured or semi structured. So we can not apply the Data mining techniques directly. Rather another discipline is evolved called Web mining that can be applied to Web data. Web mining is used to discover interest patterns which can be applied to many real world problems like improving web sites, better understanding the visitor's behavior, product recommendation etc.[10]. The above is the brief explanation of how Web usage mining is executed. Most sophisticated systems and techniques are accomplished into three main phases: such as Preprocessing, Pattern discovery and Pattern analysis. In the first phase, log files are preprocessed in order to preserve only the relevant information. In the second phase, several WUM techniques are used to identify interesting Web patterns from the preserved information. Then these Web patterns are presented for analysis in the third phase of Web usage mining. Every process can be categorized as shown in fig.1.

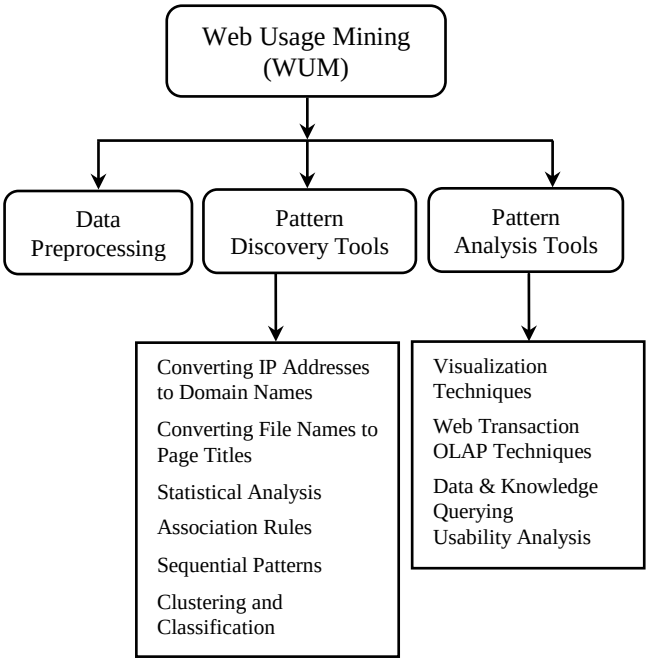

Fig. 1: Research Phases in Web Usage Mining

B. Applications of Web Usage Mining

Web usage mining has several applications and is used in the following areas:

- Web usage mining helps to determine frequent access behavior of the users, needed links can be identified to improve the overall performance of future accesses.
- Web personalization for a user can be achieved by keeping track of previously accessed pages by using WUM. These pages may also be used to identify the typical browsing behavior of a user and subsequently to predict desired pages.
- Web usage mining can be used to improve the attractiveness of a website, in terms of content and structure.
- In addition to modifications to the linkage structure, identifying common access behaviours can be used to improve the actual design of Web pages and to make other modifications to the site.
- Web patterns can be used to gather business intelligence to improve customer attraction, customer retention, sales, marketing and advertisement and cross sales etc.
- Web usage mining provides the ability to analyze massive volumes of click-stream or click flow data, integrate the data seamlessly with transaction and demographic data from offline sources and apply sophisticated analytics for Web personalization.
- Mining of Web usage patterns help in the study of how browsers are used and the user's interaction with a browser interface.
- Web usage characterization can also be viewed into navigational strategy when browsing a particular Website.
- Web usage mining focuses on techniques which are used to predict user's behavior while the user interacts with the Websites.

- Web usage mining of patterns provides a way to understanding web traffic behavior, which can be used to deal with policies on web caching, network transmission, load balancing or data distribution.
- Performance and other service quality attributes are crucial to user satisfaction and high quality performance of a web application is expected.
- Web usage and data mining is also very useful for detecting intrusion, fraud, and attempted break-ins to the system.
- Web usage mining can be used in usability studies to determine the interface quality.
- Web usage mining can be used in network traffic analysis for determining equipment requirements and data distribution in order to efficiently handle Website traffic.
- Web usage mining can be used in determination of common behaviors or traits of users who perform certain actions.
- Web usage mining can be used in Counter terrorism and fraud detection and detection of unusual accesses to secure data.
- It can be used in e-Learning, e-Business, e-Commerce, e-CRM, e-Services, e- Education, e-News papers, e-Government and Digital Libraries.

III. OVERVIEW OF WEB LOGS

The quality of the web patterns discovered in Web usage mining process highly depends on the quality of the data used in the mining processes. Web Log files record activity information when a web user submits a request to a web Server. The main source of raw data is the Web access log which is known as Log file. The Log file can be analyzed over a time period. The time period can be specified on hourly, daily, weekly and monthly basis. The following information can be gathered from the Log files of a Website:

- General Summary on number of Visitors and accessibility.
- Total Hits on number of pages/files accessed or attempted to be accessed.
- The source Websites of the Visitors.
- The Browser used by the Visitor to access the Website.
- Error Reports for identifying the problems and solving them.
- Report based on File size, File type and directory/subdirectory visited.

A. Types of Web Server Logs

Web server logs are plain text (ASCII) files and are independent from the Server. There are some distinctions between server software, but traditionally there are four types of web server logs, which are shown in fig. 2.

The first two types of Web logs such as Transfer Log and Agent Log are standard. The Referrer and Agent Logs may or may not be 'Turned On' at the Server or may be added to the Transfer Log file to create an 'Extended' log file format.[12]

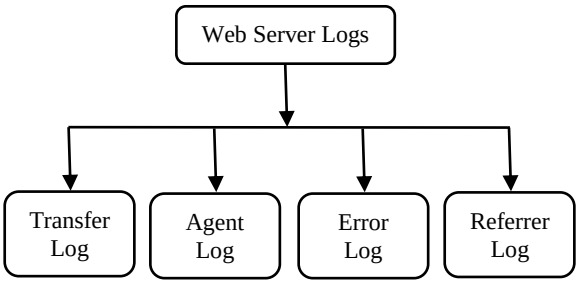

Fig. 2: Types of Web Server Logs

B. Location of the Web Logs

A web log is a file to which the web server writes information each time a user requests a website from that particular Server. If user visits many times on the website then it creates entry many times on the Server. A log file (or Web log) can be located in three different places[13]. They are as follows:

(i) **Web Servers** – These logs generally supply the most complete and accurate usage data.

(ii) **Web Proxy Servers** – A proxy server takes the HTTP requests from users and passes them to a Web server then returns to users the results passed to them by the Web server.

(iii) **Client Browsers** – Participants remotely test a Web site by downloading special software that records Web usage or by modifying the source code of an existing browser. HTTP cookies could also be used for this purpose.

Each approach suffers from some major drawbacks, which are summarized in table1 below.

Table 1

DRAWBACKS		
Server-side Logs	**Proxy-side Logs**	**Client-side Logs**
These logs contain sensitive, personal information, therefore the server owners usually keep them closed.	Proxy-server construction is a difficult task. Advanced network programming, such as TCP/IP, is required for this construction.	The design team must deploy the special software and have the end-users install it.
The logs do not record cached pages visited. The cached pages are summoned from local storage of browsers or proxy servers, not from web servers.	The proxy logger implementation in Web Quilt, a Web logging system performance declines if it is employed because each page request needs to be processed by the proxy simulator.	The technique makes it hard to achieve compatibility with a range of operating systems and web browsers.

C. Contents of a Log File

Web log file reside on the server. If user visits many times on the website then it creates entry many times on the server. The Web log file[14] has been containing the following key fields:

(i) Visiting Path– Paths which follow by the user to visit on the Website.

(ii) User Name– Identify the user through IP address which provide by ISP. It is temporary address.

(iii) Success Rate– It is user activity which is done on the Website that is number of downloads and number of copies.

(iv) Path Traversed– The path identifies who is visit on the website through user.
(v) Last visited Page– It stores the last record that is visited by the user.
(vi) URL of the Web page accessed– It may be HTML page and CGI program. This is accessed through the user.
(vii) Request Method (GET or POST): This is a method which is performing on the Website like GET and POST.

The above mentioned are the key fields present in the log file. This log file details are used in case of Web usage mining process. According to Web usage mining, it mines the highly utilized Website. The utilization may be the frequently visited Website or the Website being utilized for longer time duration. Therefore the quantitative usage of the website can be analyzed, if the log file is analyzed.

IV. LITERATURE SURVEY

Although, lot of works have been carried out in the field of Web usage mining. Web usage mining is a relative new research area, and gains more and more attentions in recent years. We will have detailed information in this section about usage mining, based on some up-to-date research works. The number of Web users doubled between the years 2005 and 2010, and was expected to surpass two billion in 2010. Early studies in 1998 and 1999 estimating the size of the web using capture or recapture methods proved that the much of the Web was not indexed by Search engines and the Web was much larger than expected. According to a 2001 study, there were a massive number, over 550 billion of documents on the Web, mostly in the invisible Web or Deep Web. A survey was done in the year 2002[30] and it has been found that among 2,024 million Web pages, the most Web pages were written in English: 56.4%; next were pages in German: 7.7%, French: 5.6%, and Japanese: 4.9%. A more recent study, which used Web searches in 75 different languages to sample the Web, determined that there were over 11.5 billion Web pages in the publicly indexable Web as of the end of January 2005. As of March 2009, the indexable web contains atleast 25.21 billion pages. On 25 July, 2008, Google software engineers Jesse Alpert and Nissan Hajaj announced that Google Search had discovered one Trillion unique URLs. As of May 2009, over 109.5 million domains operated, 74% of these were commercial or other Websites operating in the .com generic top-level domain. Statistics measuring a website's popularity are usually based either on the number of page views or on associated server 'Hits' that it receives. We have been recorded the total Websites across all the domain names from August 1995 to May 2012. The graphical presentation of the data is shown in fig. 3 below.

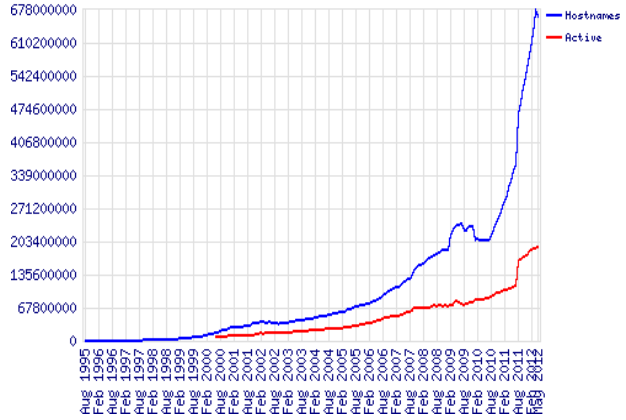

Fig.3: Total Websites across all Domains
(August 1995 - May 2012)

The existing works carried out by several researchers and it has been recorded that they mostly used the following techniques:
(a) Association Rule Mining (ARM)
(b) Clustering
(c) Classification

Table 2: The Research Work using Association Rule

Algorithms used	Researchers	Year
Maximal forward references	Ming-Syan Chen, Jong Soo Park, Philip S. Yu	1998
Markov Chains	Jianhan Zhu, Jun Hong, and John G. Hughes	2002
Improved Apriori All	WANG Tong, HE Pi-lian	2005
Fpgrowth and Prefixspan	Hengshan Wang, Cheng Yang, Hua Zeng	2006
Custom Built APRIORI Algorithm	Sandeep Singh Rawat, Lakshmi Rajamani	2010
FP Growth algorithm	Navin Kumar Tyagi, A. K. Solanki	2011

In the year 1998, Ming-Syan et al.[15] have been proposed a new data mining algorithm which involves mining path traversal patterns in a distributed information-providing environment where documents are linked together to facilitate interactive access. In the year 2002, Jianhan Zhu et al. [16] applied the Markov chains to model user navigational behavior. In this work they have investigated a technique for constructing a Markov model of a website based on past visitor behavior. In the year 2005, Wang Tong et al. [17] proposed an improved algorithm based on the original AprioriAll Algorithm. In the year 2006, Hengshan Wang et al. [18] introduced two prevalent data mining algorithms: Fpgrowth and PrefixSpan into Web usage mining. Maximum Forward Path (MFP) is also used in the Web usage mining model during sequential pattern mining along with PrefixSpan, so as to reduce the interference of 'false visit' caused by browser cache and raise the of mining frequent traversal paths. In the year 2010, Sandeep Singh Rawat et al. [19] offered a custom-built apriori algorithm that is based on the old Apriori algorithm, to find the effective pattern analysis. In the year 2011, [20] the researchers have proposed a recommendation methodology based on correlation rules. Association rules are generated from log data by using FP

Growth algorithm and then cosine measure is used for generating correlation rules.

Table 3: The Research Work using Clustering

Algorithms used	Researchers	Year
Self Organized Maps	Paola Britos, Damian Martinelli, Hernán Merlino, Ramón Garcia-Martínez	2007
Graph Partitioning	Mehrdad Jalali, Norwati Mustapha, Ali Mamat, Md. Nasir B Sulaiman	2008
Ant-based	Kobra Etminani, Mohammad-R. Akbarzadeh-T, Noorali Raeeji Yanehsari	2009
K-mean with Genetic Algorithm	N. Sujatha, K. Iyakutty	2010
EM-CFuzzy means algorithms	K.Poongothai M.Parimala, Dr. S. Sathiyabama	2011

In the year 2007, Paola Britos et al.[21] have presented the study of the use of Self Organized Maps, kind of artificial neural network, in the process of Web usage mining to detect user's patterns.

In the year 2008, Mehrdad Jalali et al.[22] proposed an approach that was based on the graph partitioning for modeling user navigation patterns. In order to mining user navigation patterns, they establish an undirected graph based on connectivity between each pair of the web pages and also proposed novel formula for assigning weights to edges of the graph. In the year 2009[23], Kobra Etminani et al., applied ant-based clustering algorithm to pre-processed logs to extract frequent patterns for pattern discovery and then it is displayed in an interpretable format. In the year 2010, N. Sujatha et al., [24] have proposed a new framework to improve the web sessions' cluster quality from k-means clustering using Genetic Algorithm (GA). In the year 2011, [25] the web usage mining framework presented that evaluates the performance of expectation-maximization (EM) and CFuzzy means cluster algorithms. The proposed Miner framework is an initial effort to patch up some of the weaknesses of the conventional Web log file analyzers. The experimentation on the K-means clustering is also conducted.

Table 4: The Research Work using Classification

Algorithms used	Researchers	Year
Naive Bayesian	Mahdi Khosravi, Mohammad J. Tarokh	2010

In the year 2010, Mahdi Khosravi et al., [26] proposed a novel approach for dynamic mining of users' interest navigation patterns, using Naive Bayesian method.

V. METHODOLOGY

A. Comparative Study of different Algorithms:

In our work we offer a comparative study of different algorithms which have been used by most of researchers in their work. The Association Rule Mining, Clustering and Classification techniques have been discussed on their characteristics. The summary of these algorithms are also presented here.

a. Comparison of ARM Algorithms

The merits and demerits of these Association Rule Mining Algorithms are summarized in table 5.

Table 5: Comparison of Association Rule Algorithms

Algorithm	Merits	Demerits
Apriori	Same as for sequential pattern discovery	---
Market Basket Analysis	Produces simple and easy to understand results	Only works for discrete data values otherwise binning is necessary. Works best when web pages occur in roughly the same number of user Sessions.

According to the above table, it has been found that Market Basket Analysis is not as effective when Web pages do not occur in approximately the same number of user sessions. Then the Apriori algorithm will be used as the association rule algorithm for analyzing Web usage patterns.

b. Comparison of Clustering Algorithms

The merits and demerits of these clustering algorithms are summarized in table 6.

Table 6: Comparison of Clustering Algorithms

Algorithm	Merits	Demerits
k-Means	Relatively scaleable	Mean needs to be defined (Smith and Ng 2003). The number of clusters needs to be specified. Noisy data and outliers cause problems.
Automatic Cluster Detection	Works well with categorical, numerical and textual data (Brooks 1997).	Can be difficult to choose the correct distance measures and weights. Can be hard to interpret the resulting clusters.

In this context, since automatic cluster detection works well with textual data and is therefore suited to Web usage data, it can used for cluster analysis in the proposed model.

c. Comparison of Classification Algorithms

The merits and demerits of these classification algorithms are mentioned in table 7 below.

Table7: Comparison of Classification Algorithms

Algorithm	Merits	Demerits
Decision Trees	Can handle raw data. Requires little preprocessing. Produces easy to understand rules. Provide clear indication of which fields are most important for prediction and classification. Can handle large number of fields.	Mistake at higher level causes all lower levels to be incorrect. Does not easily handle non-numeric data. Computationally expensive.

Naïve Bayesian Classifiers	Works well with numeric data. Easy to use. Requires only one pass through the data. Works with partial data.	Performs very poorly in cases of high correlation. Limited to discrete variables.
k-Nearest Neighbour	Can utilize entire data source rather than require sampling for training. Good for discovering clusters.	Requires large amounts of memory. May be overly sensitive to closely matching records.

Since these algorithms do not easily handle non-numeric data such as the Log file data, enumeration of the Log files are required before analysis can be performed. It has been decided that classification is beyond the scope of this work, due to the fact that additional information regarding the Web pages as well as the users of the Web site is required.

VI. EXPERIMENTAL EVALUATION

A. Web Usage Mining Tools

In order to obtain a better understanding of Web usage mining, several related Web mining tools are investigated in this study. Many web traffic analysis tools, such as WebTrends and WebMiner, are available for generating Web usage statistics. We have been used one of three related Web mining tools. These Web Mining tools are: Absolute Log Analyzer, WebLog Expert, and 123LogAnalyser. The criteria that were used when evaluating these tools correspond to the research phases of Web usage mining (fig.1). This paper aims to determine for each Web mining tool, the type of data storage that takes place in the preprocessing, what Web mining algorithms are used in the pattern discovery and how these patterns are visualised in the pattern analysis.

A.1 Absolute Log Analyzer Tool

Absolute Log Analyzer [27] is a client-based log file analysis software tool, designed for Web traffic analysis. Firstly, log files need to be added to the analysis and the results are then displayed. Apart from the graphical user interface (GUI), Absolute Log Analyzer also has a command line interface (CLI).

a. Data Storage

Absolute Log Analyzer allows log files to be downloaded via FTP. The analyser can recognize the majority of log files format (Microsoft IIS and Apache) automatically. It also has the facility to manually specify your own format for non standard log files. It will analyse compressed log files (.gz and .zip) and can recompress them to minimize drive space usage. This tool imports data into the highly optimised proprietary database. This allows the user to incrementally update the statistics as new log files become available and makes it simple to zoom in on a particular quarter, month, week, or day and even view all of these statistics in the same table, so that any trends can be evaluated.

b. WUM Algorithms

The screen shot of the Web mining tool Absolute Log Analyzer is shown in fig. 4 below that displays the settings for the analysis. These settings are used to tailor various aspects of the analysis and are categorized by the tabs at the top of the window. It is not apparent, however, which WUM Algorithms are used for the analysis, nor is there any way of selecting alternate algorithms. Only descriptive statistics are provided.

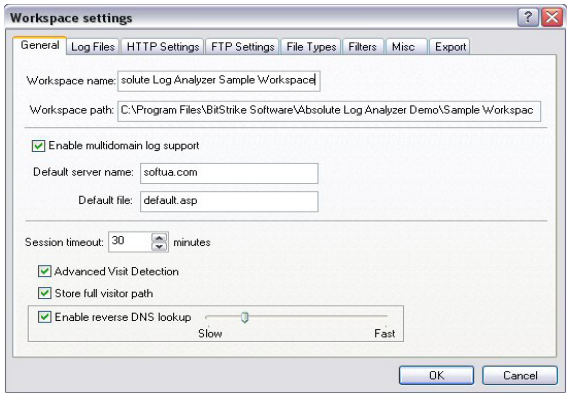

Fig. 4: Options Window of Absolute Log Analyzer

c. Visualization of Results

As discussed previously, the report generated by this tool Absolute Log Analyzer is displayed in the main screen as shown in fig. 5 below. It does not provide an option to export the full report or subsections of it into HTML format. The menu on the left hand side displays the available statistics. The textual results are shown in the right hand side of the window and where applicable, a graphical representation of this analysis is shown in the window at the bottom of the screen.

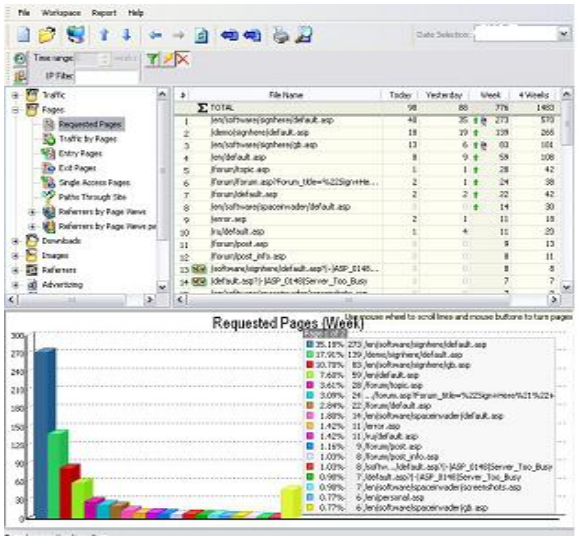

Fig. 5: Report by Absolute Log Analyzer

The information provided in the report is summarized in table 8 below.

Table 8: Analysis Report from Absolute Log Analyzer

Information Provided	Description
Traffic	Reports in this group describes general site information and includes summary report, list of visits during particular day and three types of information about requests to the server, used bandwidth, visits, page hits, downloads, images, bots and spiders, web server errors.
Pages	This group contains information about page views on the site. Particular reports are most visited pages, entry pages, exit pages, single access pages and paths through site.

Downloads	This group is dedicated to resources with type 'Download'. Here reports about downloaded files, used bandwidth, referrers and referrer relevancy analysis can be seen.
Images	Here information about banners or other important image files is seen
Referrers	Contains complete information about site's referrers such as how many visitors comes from a particular page, server or search engine, which search words and phrases have been used to find the site and so on.
Audience	Provides more information about visitors, including where they reside, which operating system uses and more.
Bots and Spiders	Information about bots, spiders, crawlers and other non-human activity on the site.
Technical Information	Broken link reports and page not found errors can be appeared here.

A.2 WebLog Expert Tool

WebLog Expert is a fast and powerful Web log Analyzer Web mining tool [28]. This software tool assists to reveal important statistics regarding a Web site's usage like: activity of visitors, access statistics, paths through the website, visitors' browsers, and much more.

a. Data Storage

WebLog Expert is also a powerful Web mining tool that supports the W3C Extended log format that is the default log format of Microsoft IIS 4.0/.05/6.0/7.0. This software tool also supports the Combined and Common log formats of Apache Web server. It supports compressed log files (.gz, .bz2 and .zip) and can automatically detect the log file format. If necessary, log files can also be downloaded via FTP or HTTP. Analysis is performed directly on the web log files and no separate data warehouse is required.

b. WUM Algorithms

The options window of WebLog Expert software tool is shown in a fig. 6 below. The GUI interface of this tool displays all analysis options, reporting settings and log file download settings on this screen. It is possible to schedule an analysis to take place automatically. It is, however, not apparent which WUM algorithms are used for this analysis and only descriptive statistics are provided.

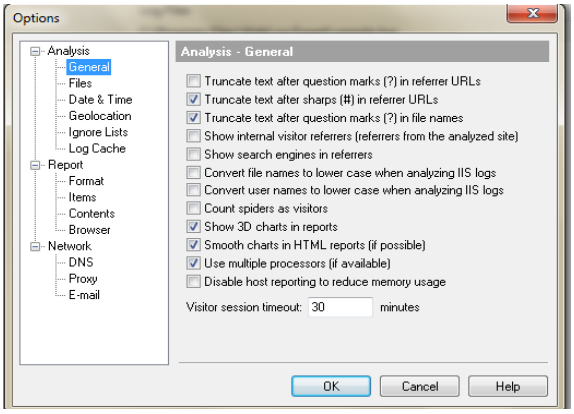

Fig. 6: Options Window of WebLog Expert

Once the log files have been selected there is an include/exclude filter, which allows the user to select what information should be included or excluded from the analysis.

c. Visualization of Results

WebLog Expert tool generates an easy-to-read HTML report. This HTML report contains various categories which can be navigated from the menu on the left hand side. The textual results together with simple bar and line graphs are then displayed on the right hand side as is shown in fig. 7 below. If an analysis has been scheduled, the generated report can be automatically e-mailed to the Web designer, if required.

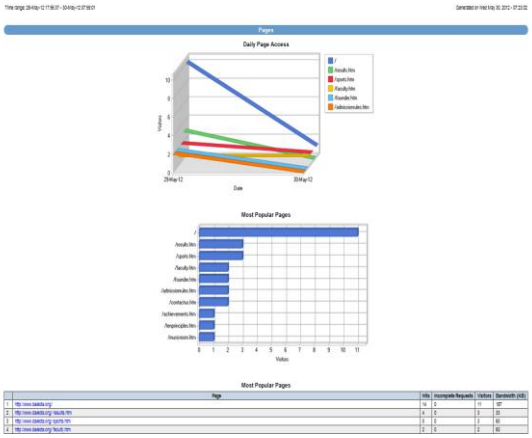

Fig. 7: Report by WebLog Expert

The information provided in the HTML report is summarized in table 9 below.

Information Provided	Description
General statistics	This shows total and average hits, total and average page views, total and average number of visitors, total and average bandwidth
Activity statistics	daily, by hours of the day, by days of the week and by months
Access statistics	This displays statistics for pages, files, images, directories, entry pages, exit pages, paths through the site and file types
Information about visitors	hosts, top-level domains, countries, states, cities, organizations, authenticated users
Referrers	referring sites, URLs, search engines (including information about search phrases and keywords)
Browsers, Operating systems and Spiders statistics	most frequently used browsers and operating systems as well as frequently detected spiders
Information about errors	error types, detailed 404 error information
Tracked files statistics	This displays the activity and referrers

Table 9: Analysis Report from WebLog Expert

A.3 123Log Analyzer

123LogAnalyzer is a popular and powerful tool developed by ZY Computing Inc. in 2003[29]. It is a web traffic analyzing tool which is the fastest web log analyzing tool in the market. It is a Windows-based program which can read the major log file formats from both UNIX and Windows platforms. It is simple and its intuitive interface requires no technical knowledge. It can analyze a log file at 650MB per minute (40,000 lines per second).

On a 500 Mhz PIII Computer running Windows 2000 it can analyze a 625MB log file in only 54 seconds. 123LogAnalyzer offers deeper research capabilities and more information than other analyzing tools.

a. Data Storage

One useful feature of 123Log Analyzer is the program's ability to analyse log file archives (such as .zip or .gz) without the need to extract the files to the client machine first. Retrieving and analysing compressed logs from a remote location can also save some download time and hard drive space on the client machine. 123LogAnalyzer does not, however, allow multiple log files to be in the same archive. In addition to allowing files to be manually added for analysis, 123LogAnalyzer also allows the files to be downloaded directly from a remote location via FTP or HTTP. The log file types that are accepted as input are .log and .txt.

123Log Analyzer performs the analysis directly on the log files without duplicating the data. For this reason, no separate data warehouse is required.

b. WUM Algorithms

Once log files have been added for analysis, various filters can be applied in order to perform an in-depth and precise analysis of the data. The fig. 8 below shows the filtering options available in this tool.

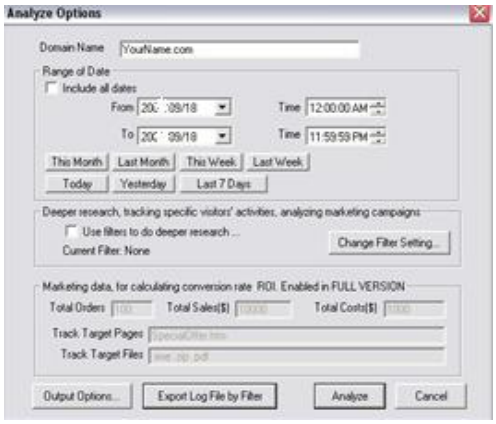

Fig. 8: Options Window of 123LogAnalyzer

These filters can enable the user to fine-tune the results. For example, the user can determine how many users visited a particular web page and also what other web pages they visited, what their browsing sequences are, which files they downloaded, which web site they were referred from, and what keywords they used to find the web site being analysed. Combinations of these filters can also be used. It is not evident, however, from the investigation or from the documented reports, exactly what WUM algorithms are used for this analysis and only descriptive statistics are provided.

c. Visualization of Results

Once the analysis has been performed using this Web mining tool, an HTML report is generated and is displayed with the help of the Web browser. This HTML report is shown in fig. 9 given below. The information is categorized in the left hand column and the results are displayed on the right hand side.

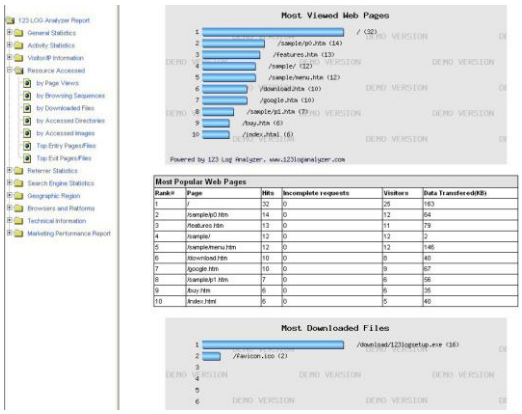

Fig. 9: Report by 123LogAnalyzer

Where applicable, simple 2D bar and line graphs of the generated report are also displayed. The information provided in the HTML report is summarized in table 10 below.

Table 10: Analysis Report from 123LogAnalyser

Information Provided	Description
General statistics	This shows a daily visit report graph and a statistics table for the time period entered for the range of dates in the analyse window. It also provides a summary of the activity statistics.
Activity statistics	This displays the number of visitors, unique IP's, amount of bandwidth used, and the number of hits, broken down by Time Increment, Day of the Week, and Hour of the Day for the time period entered. Analyses by Pageview per Day and by Visitor Stay Length can also be viewed.
Visitor/IP Information	This displays the IP addresses, Domain Name, Country, Time of Last Access, and IP Ownership information for the site's visitors, broken down by Access Time, Hits, Bandwidth, Stay Length, and Authenticated Visitors.
Resource Accessed	This provides web pages viewed, files downloaded, directories that were accessed, and images that were accessed during the time period, broken down by Page Views, Browsing Sequences, Downloaded Files, Accessed Directories, Accessed Images, Top Entry Pages and Files, and Top Exit Pages and Files.
Referrer Statistics	This report shows which Domains and URLs the visitors come from according to Referring Domains and Referring URLs.
Search Engine Statistics	This report displays the search engines that referred visitors to the site, the phrases and keywords visitors searched for broken down by Top Search Engines, Keywords, and Each Search Engine.
Geographic Region	This report displays a Most Active Countries graph and table showing which Countries the visitors come from during the time period.
Browsers and Platforms	This report shows which browsers and platforms were used by visitors who visited during the time period.
Technical Information	An explanation of errors encountered is provided.
Marketing Performance Report	The Marketing Report helps analyse the online marketing campaign. 123LogAnalyzer tracks costs, sales, and profits by visitors, by bandwidth, and by page views and downloads. Information regarding marketing costs and tracked pages needs to be entered.

B. COMPARISON OF WUM TOOLS

The most popular and powerful three web mining tools are applied to analyze the web access logs from a website of an educational institute. These tools provide descriptive statistics regarding the activity of the server in terms of files requested, referring websites and peak traffic times.

They also provide useful bar and line graphs for representing the statistics generated and information regarding frequent navigation paths through the website however they do not sufficiently illustrate browsing characteristics such as similarities between the multiple sessions for a single user or files accessed frequently between various sessions. Finally, these software tools do not provide a graphical representation of the users' navigation paths. The textual representation of these paths provided by these tools is difficult to interpret since these are simply combinations of the URLs visited. Our experience has shown that there is no single best powerful tool for all web usage mining applications.

VII. CONCLUSION

Web mining software tools are very costly. Selection of wrong tool is expensive both in terms of money and loss of time. The work presented in this paper provides a method for evaluating and selecting an appropriate tool. This paper has evaluated three related web mining tools by comparing the WUM algorithms as well as the visualization results presented by each tool. In this work, the three Web mining tools are investigated such as: Absolute Log Analyzer, WebLog Expert and 123Log Analyzer. These tools need to have specific knowledge about the particular problem domain to do any more than filtering based on statistical attributes of the discovered rules or patterns. Existing WUM tools however, do not indicate which Web usage mining algorithms are used or provide effective graphical visualizations of the results obtained. The information is gathered on a daily basis and continues to be analyzed consistently. More research needs to be done in Computer Security, Web Intelligence, Intelligent Learning, Bioinformatics, Healthcare and Telecommunications by using Web usage mining.

ACKNOWLEDGMENT

The author express deep gratitude to Dr. P.C. Gupta, Professor & HOD Computer Science, School of Engineering and Technology, Jaipur National University, Jaipur, India for the encouragement and extensive support in preparing and publishing of this paper.

REFERENCES

1. The W3C Technology Stack; World Wide Web Consortium; Retrieved April 21, 2012
2. http://whatis.techtarget.com on May 24, 2012
3. Zaiane O.; Conference Tutorial Notes: Web Mining Concepts, Practices and Research; In Proc. SDBD 2000
4. Oren Etzioni; The World Wide Web: Quagmire or gold mine. Communications of the ACM, 39(11); 65-68, 1996
5. R. Kosala, and et al.; Web mining Research: A Survey
6. S.K. Madria et al.; Research issues in Web data mining; In Proceedings of Data Warehousing and Knowledge Discovery, First International Conference, DaWaK'99, pages 303-312, 1999
7. R. Cooley, Web Usage Mining: Discovery and Application of Interesting Patterns from Web data, University of Minnesota, May 2000
8. Piatetsky Shapiro g., and et al.; Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996
9. Liu B.: Web Data Mining: Exploring Hyperlinks, Contents and Usage Data, Springer Berlin Heidelberg New York, 2006
10. S.K. Pani, and et al.; Web Usage Mining: A Survey on Pattern Extraction from Web Logs; International Journal of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1, 2011
11. Cooley R., et al.; Web mining: Information and Pattern Discovery on the World Wide Web. A survey paper; In: Proc. ICTAI97, 1997
12. L.K. Joshila Grace, and et al.; 'Web Log Data Analysis and Mining' in Proc CCSIT-2011, Springer CCIS, Vol 133, (Jan 2011), pp 459-469
13. K. R. 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, pp. 327-332, 2009
14. Ratnesh Kumar Jain, and et al.; 'Efficient Web Log Mining using Doubly Linked Tree', International Journal of Computer Science and Information Security, IJCSIS, Vol. 3, July 2009
15. Ming-Syan Chen, and et al.; 'Efficient Data Mining for Path Traversal Patterns', IEEE Transactions on Knowledge and Data Engineering, Vol. 10, No. 2, March/April 1998
16. Jianhan Zhu, and et al.; 'Using Markov Chains for Link Prediction in Adaptive Web Sites', Soft-Ware 2002, LNCS 2311, pp. 60–73, 2002
17. WANG Tong, and HE Pi-lian, 'Web Log Mining by an Improved Apriori All Algorithm', World Academy of Science, Engineering and Technology, Vol. 4, 2005
18. Hengshan Wang, and et al.; 'Design and Implementation of a Web Usage Mining Model Based on Fpgrowth and Prefixspan'; Communications of the IIMA, Vol. 6, Issue 2, 2006
19. Sandeep Singh Rawat, and et al.; 'Discovering Potential User Browsing Behaviors Using Custom-Built Apriori Algorithm', International Journal of Computer Science & Information Technology (IJCSIT) Vol.2, No.4, August 2010
20. Navin Kumar Tyagi, and et. Al; 'Analysis of Server Log by Web Usage Mining for Website Improvement'; IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No. 8, July 2011
21. Paola Britos, and et al.; 'Web Usage Mining Using Self Organized Maps', International Journal of Computer Science and Network Security, Vol.7 No.6, June 2007
22. Mehrdad Jalali, and et al.; 'Web User Navigation Pattern Mining Approach Based on Graph Partitioning Algorithm', Journal of Theoretical and Applied Information Technology, Pakistan
23. Kobra Etminani, and et al.; 'Web Usage Mining: Users' navigational patterns extraction from web logs using Ant-based Clustering Method', IFSA-EUSFLAT 2009
24. N. Sujatha, K. Iyakutty, 'Refinement of Web usage Data Clustering from K-means with Genetic Algorithm', European Journal of Scientific Research ISSN 1450-216X Vol.42 No.3 (2010), pp.464-476
25. K.Poongothai et al., 'Efficient Web Usage Mining with Clustering'; IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011
26. Mahdi Khosravi et al., 'Dynamic Mining of Users Interest Navigation Patterns Using Naive Bayesian Method', 978-1-4244-8230-6/10/©2010 IEEE
27. http://www.bitstrike.com (Last Accessed: 24 May 2012) Bitstrike Software (2004): Absolute Log Analyzer.
28. http://www.weblogexpert.com (Last Accessed: 24 May 2012) Alentum Software Inc (2004): WebLog Expert.
29. http://www.123loganalyzer.com (Last Accessed: 24 May 2012) Zy Computing Inc (2003): 123 Log Analyzer. San Jose, USA.
30. http://en.wikipedia.org seen on May 2012.

AUTHORS PROFILE

Arvind K. Sharma, obtained his Master's Degree in Computer Application from JNRV University, Udaipur and M.Phil Computer Science from Alagappa University. He is currently pursuing Ph.D. Computer Science from Jaipur National University Jaipur, Rajasthan-India. He also has publications in National and International Journals. His research interest lies in the area of Data Mining, Web Mining & Web Applications.