An Investigative Approach of Web Data Mining Tools for Web Analytics

[1]Dr. Arvind K Sharma, [2]Dr. Anubhav Kumar
[1]Research Supervisor, Career Point University, Kota, Rajasthan, India
[2]Dept. of CSE, IET College, Alwar, Rajasthan, India

Abstract

Web mining is the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents usage of web sites. The web is the interesting area of research. With the help of Web mining, the user obtains the required information accurately. Web mining is categorized into three types: web content mining, web structure mining and web usage mining. In this paper the taxonomy of web mining and an investigative approach of several web data mining tools for web analytics are presented.

Keywords

Web Mining, Web Content Mining, Web Structure Mining, Web Usage Mining, Web Data Mining Tools

I. Introduction

The concept of WWW was given in 1989 by Tim Berners-Lee while at CERN (the European Laboratory for Particle Physics). Today, WWW is a popular and interactive medium to interchange information. 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 [1]. 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 [2]. 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. No one 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 [3]. 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 uses the Client/Server model and the World Wide Web's Hypertext Transfer Protocol (HTTP), serves the files that form web pages to web users [4]. 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 [5]. 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, and Success Rate. Web mining studies, analyzes and reveals useful information from the Web [6]. 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 interest. The World Wide Web (WWW) is increasingly growing with the information transaction volume from Web servers and the number of requests from Web users in Internet. Analyzing of web server access logs is one of the application areas of web mining. With the rapid growth of the World Wide Web (WWW), it becomes more important to find the useful information from these huge amounts of data. The Web also contains a rich and dynamic collection of hyperlink information and Web page access and usage information [12]. This is due to the exponential growth of the World Wide Web and its architecture and also due to the increase of its importance over the people's life.

The rest of paper is organized as follows: Section II describes web mining and its different categories. Section III provides an approach of web data mining tools along with their characteristics. In section IV we conclude the paper with summary. Finally, in the last section references are mentioned.

II. Web Mining

Web mining is the term of applying data mining techniques to automatically discover and extract useful information from the World Wide Web documents and services [8]. Web mining is the application of data mining techniques to extract knowledge from Web data, including Web documents, hyperlinks between documents, usage logs of web sites, etc. Web mining can be broadly divided into three distinct categories, according to the kinds of data to be mined that are web content mining, web structure mining and web usage mining.

A. Taxonomy of Web Mining

In this section we present taxonomy of web mining. The web mining is the use of data mining techniques to automatically discover and extract information from web documents and services [9] in which at least one of structure or usage (web log) data is used in the mining process.

Web mining can be categorized into three distinct categories. This taxonomy is shown in fig. 1.

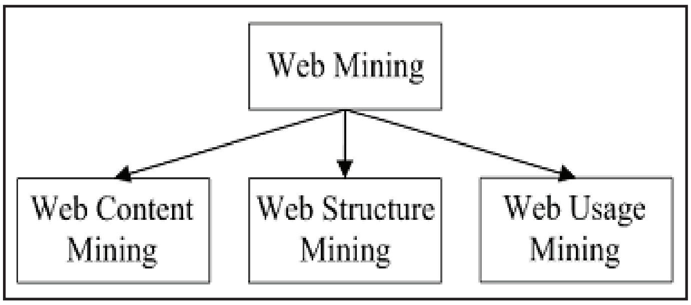


Fig. 1: Taxonomy of Web Mining

1. Web Content Mining

Web Content Mining is the process of picking up useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may contain text, images, audio, video or structured records such as lists and tables. Research activities in this field also involve using techniques from other disciplines [10] such as Information Retrieval (IR) and Natural Language Processing (NLP).

2. Web Structure Mining

Web Structure Mining tries to identify the structure of hyperlink in html documents and deduce knowledge [11]. It is a process of picking up information from linkage of web pages. It operates on the web's hyperlink structure. Web structure mining is also a process of using graph theory to analyse the node and connection structure of a web site. The structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting between two related pages. In addition, the content within a web page can also be organized in a tree-structured format, based on the various Hyper Text Markup Language (HTML) and eXtensible Markup Language (XML) tags within the page.

3. Web Usage Mining

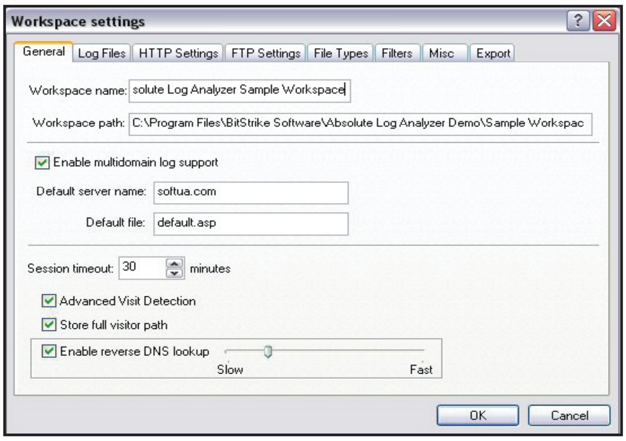
Web usage mining is also known as Web log mining. It is a process of picking up information from user, how to use web sites. It is an application of data mining techniques to discover interesting usage patterns from web data, in order to understand and better serve the needs of web based applications. Usage data captures the identity or origin of web users along with their browsing behavior at a website [7]. Some of the typical usage data collected at a web site includes IP addresses, page references and access time of the users. The web usage data contains the data from Web server access logs, Proxy server logs, Browser logs, User profiles, Registration data, User sessions or transactions, Cookies, User queries, Bookmark data, Mouse clicks and Scrolls and any other data as the results of interactions.

III. Web Data Mining Tools

In this section, we are going to present several web data mining tools in which some of the tools are open source softwares and freely available over the internet. These tools use the descriptive statistics method; due to problems installing other programmes that work with the other methods programmes from this category were chosen. The choice of programmes within the category was based on the fact that for comparison reasons, we need a commercial programme, a freeware and a shareware. Many web traffic analysis tools, such as WebTrends and WebMiner are available for generating web log statistics. We will be using one of the web mining tools in our upcoming research work. Some of them are discussed.

A. Absolute Log Analyzer

Absolute Log Analyzer [13] 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). It allows log files to be downloaded via FTP. The analyser can recognize the majority of log files format. 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. The screenshot of the Web mining tool Absolute Log Analyzer is shown in fig. 2. It displays workspace settings and analysis. These settings are used to tailor various aspects of the analysis and are categorized by the tabs at the top of the window.



Workspace name: solute Log Analyzer Sample Workspace

Workspace path: C:\Program Files\BitStrike Software\Absolute Log Analyzer Demo\Sample Workspac

☑ Enable multidomain log support

Default server name: softua.com

Default file: default.asp

Session timeout: 30 minutes

☑ Advanced Visit Detection

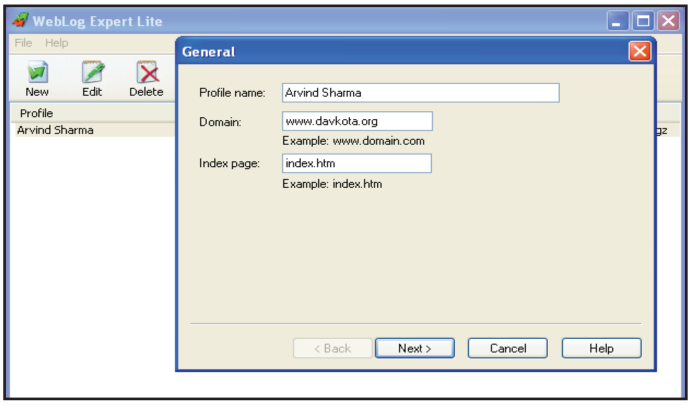
☑ Store full visitor path

☑ Enable reverse DNS lookup

Fig. 2: Options Window of Absolute Log Analyzer

B. WebLog Expert Lite

WebLog Expert Lite is a fast and powerful web mining tool. It helps 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[19]. It supports the W3C Extended log format that is the default log format of Microsoft IIS 4.0/.05/6.0/7.0. This 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[14]. If necessary, log files can also be downloaded via FTP or HTTP. GUI Interface of WebLog Expert is shown in fig. 3.



Arvind Sharma

Profile name: Arvind Sharma

Domain: www.davkota.org

Example: www.domain.com

Index page: index.htm

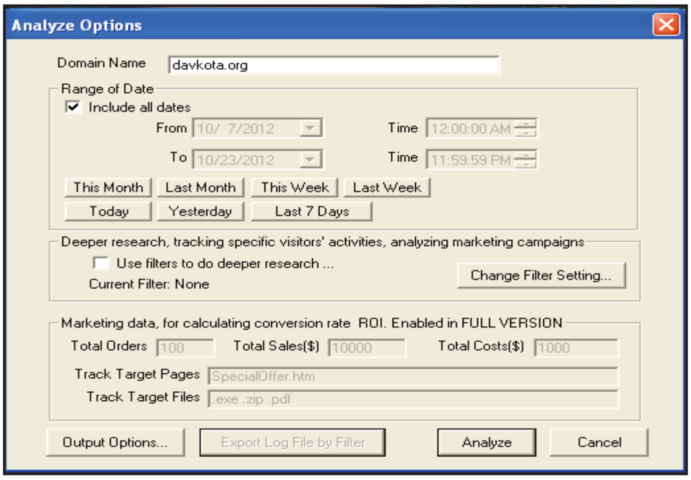
Example: index.htm

Fig. 3: User Interface of WebLog Expert Lite

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. 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 [25]. It generates an easy-to-read HTML report.

C. 123Log Analyzer

123LogAnalyzer is a popular and powerful tool developed by ZY Computing Inc. in 2003 [15]. 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 P-III 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. 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 [15]. The log file types that are accepted as input are .log and .txt. It performs the analysis directly on the log files without duplicating the data. For this reason, no separate data warehouse is required. 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. Fig. 4 shows the filtering options available in this tool.


Analyze Options

Domain Name davkota.org

Range of Date

☑ Include all dates

From 10/ 7/2012 Time 12:00:00 AM

To 10/23/2012 Time 11:59:59 PM

Deeper research, tracking specific visitors' activities, analyzing marketing campaigns

□ Use filters to do deeper research ...

Current Filter: None

Marketing data, for calculating conversion rate ROI. Enabled in FULL VERSION

Total Orders 100 Total Sales($) 10000 Total Costs($) 1000

Track Target Pages SpecialOffer.htm

Track Target Files .exe .zip .pdf

Fig. 4: Options Window of 123LogAnalyzer

D. FastStats 2.73

This web analyzing tool is a shareware program-in fact, the contribution is very low: the home page of the company is http://www.mach5.com. It can go through the entries in our log file quickly and generate statistics and reports. A screenshot of the program is shown in the fig. 5.

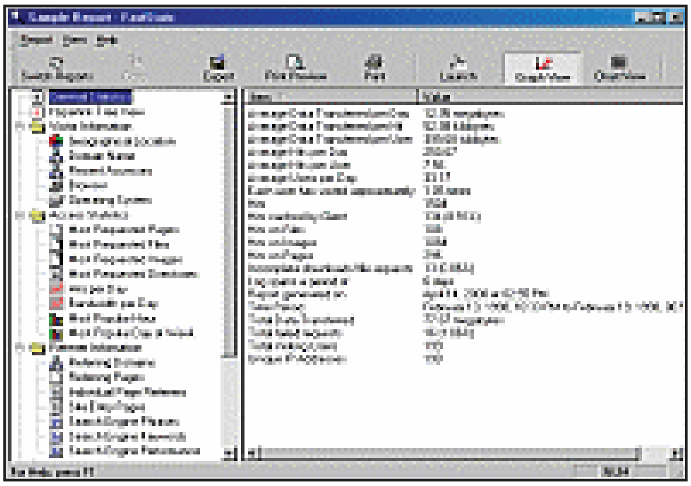


Fig. 5: A Screenshot of FastStats 2.73 Tool
(Adopted from the Tucows Website, www.tucows.com)

The tool starts with a screen, which contains the reports that have been recently accessed; the user has a choice between adding, editing, deleting, generating or copying a report- the option of cancelling- is also available. As far as the reports that are already on the list, their properties are shown as well. These properties involve the location of the log file, the existence of any filters-specific users to look for, whether the DNS retrieval option from IP addresses is enabled or not, and whether the user wants path analysis to be performed. Since we want a new report, we choose the option Add Report (we need to delete the sample report that already exists). We then need to specify whether the log files are stored locally, that is on a PC drive, on an FTP server or on a web server.

E. DTREG

It is a commercial software tool for predictive modelling and forecasting offered are based on decision trees, SVM, Neural Network and Gene Expression programs [16]. For clustering, the property page contains options that ask the user for the type of model to be built (e.g. K-means). It can also build model with a varying number of clusters or fixed number of clusters. We can also specify the minimum number of clusters to be tried. If the user wishes, then it has options for selecting some restricted number of data rows to be used during the search process. Once the optimal size is found, the final model will be built using all data rows. It has parameters like cross validate folds, Hold out sample percentage, usage of training data which evaluate the accuracy of the model for each step. It provides standardization and estimation of importance of predictor values. We can also select the type of validation which DTREG should use to test the model.

F. Cluster3

Cluster3 is an open source clustering software available here contains clustering routines that can be used to analyze gene expression data. Routines for partitional methods like k-means, k-medians as well as hierarchical (pairwise simple, complete, average, and centroid linkage) methods are covered.It also includes 2D self-organizing maps. The routines are available in the form of a C clustering library, a module of perl, an extension module to Python, as well as an enhanced version of Cluster, which was originally developed by Michael Eisen of Berkeley Lab. The C clustering library and the associated extension module for Python was released under the Python license. The Perl module was released under the Artistic License [17].

G. CLUTO

CLUTO is a software package for clustering low- and high-dimensional datasets and for analyzing the characteristics of the various clusters. CLUTO is well suited for clustering data sets arising in many diverse application areas including information retrieval, customer purchasing transactions, web, GIS, science, and biology. CLUTO[18] provides three different classes of clustering algorithms that are based on the partitional, agglomerative clustering, and graph partitioning methods. An important feature of most of CLUTO's clustering algorithms is that they treat the clustering problem as an optimization process which seeks to maximize or minimize a particular clustering criterion function defined either globally or locally over the entire clustering solution space. CLUTO has a total of seven different criterion functions that can be used to drive both partitional and agglomerative clustering algorithms which are described and analyzed in [19]. CLUTO's distribution consists of both stand-alone programs (vcluster and scluster) for clustering and analyzing these clusters, as well as, a library via which an application program can access directly the various clustering and analysis algorithms implemented in CLUTO. Its different versions are available: gCLUTO, wCLUTO.

H. Clustan

Clustan is an integrated collection of procedures for performing cluster analysis [20]. It helps in designing software for cluster analysis, data mining, market segmentation, and decision trees [21].

I. Octave

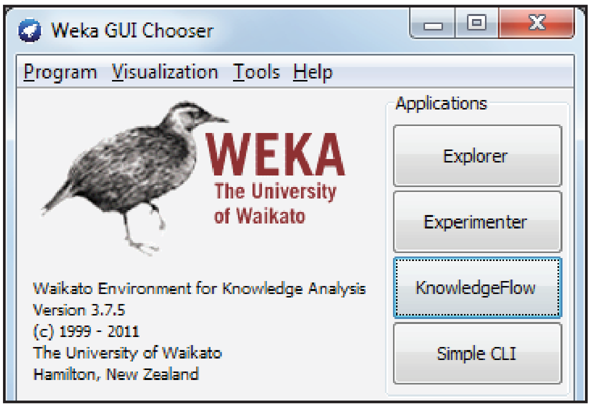
It is free software similar to Matlab and has details in [22].

J. SPAETH2

It is a collection of Fortran 90 routines for analyzing data by grouping them into clusters [23].

K. WEKA

WEKA stands for Waikato Environment for knowledge analysis. Weka is software available for free used for machine learning [24]. It is coded in Java and is developed by the University of Waikato, New Zealand. Weka workbench includes set of visualization tools and algorithms which is applied for better decision making through data analysis and predictive modeling. It also has a GUI (graphical user interface) for ease of use. It is developed in Java so is portable across platforms Weka has many applications and is used widely for research and educational purposes. Data mining functions can be done by Weka involves classification, clustering, feature selection, data pre-processing, regression and visualization. Weka GUI Interface screen is shown in fig. 6.



Waikato Environment for Knowledge Analysis
Version 3.7.5
(c) 1999 - 2011
The University of Waikato
Hamilton, New Zealand

Fig. 6: GUI Interface of WEKA

This is a Weka GUI Chooser. It provides four interfaces:

- **Explorer:** It is used for exploring the data with Weka by providing access to all the facilities by the use of menus and forms.
- **Experimenter:** Weka Experimenter allows you to create, analyse, modify and run large scale experiments. It can be used to answer question such as out of many schemes which is better.
- **Knowledge Flow:** It has the same function as that of explorer. It supports incremental learning. It handles data on incremental basis. It uses incremental algorithms to process data.
- **Simple CLI:** CLI stands for command line interface. It just provides all the functionality through command line interface.

L. Screen-scaper [26]

Screen-scraping is a tool for extracting/mining information from web sites. It can be used for searching a database, SQL server or SQL database, which interfaces with the software, to achieve the content mining requirements. The programming languages like Java, .NET, PHP, Visual Basic and Active Server Pages (ASP) can also be used to access screen scraper.

M. Automation Anywhere 6.1 (AA)

AA is a Web data extraction tool used for retrieving web data, screen scrape from Web pages or use it for Web mining[27].

N. Web Info Extractor (WIE)

This is a tool for data mining, extracting Web content, and Web content analysis [28]. WIE can extract structured or unstructured data from Web page, reform into local file or save to database, place into Web server.

O. Web Content Extractor (WCE) [29]

WCE is a powerful and easy to use data extraction tool for Web scraping, data mining or data extraction from the Internet. It offers a friendly, wizard-driven interface that will help through the process of building a data extraction pattern and creating crawling rules in a simple point-and click manner. This tool allows users to extract data from various websites such as online stores, online auctions, shopping sites, real estate sites, financial sites, business directories, etc. The extracted data can be exported to a variety of formats, including Microsoft Excel (CSV), Access, TXT, HTML, XML, SQL script, MySQL script and to any ODBC data source.

P. Mozenda [30]

This tool enables users to extract and manage Web data. Users can setup agents that routinely extract, store, and publish data to multiple destinations. Once information is in Mozenda systems, users can format, repurpose, and mashup the data to be used in other applications or as intelligence.

IV. Conclusion

It is concluded that web mining is used to retrieve online data. Data is stored in server database in web mining and it can handle multiple transactions at the same time. Data can be discovered and extracted from multiple locations of the world by sitting at one location and is able to provide desired information at the time of requirement. The main uses of web mining are to gather, categorize, and organize best possible information available on the Web to the user requesting the information. The web data mining tools are imperative to scan hypertext documents, images, and text provided on the web pages.

In this paper an investigative approach of different web data mining tools is shown.

References

[1] Piatetsky Shapiro G. et al., "Advances in Knowledge Discovery and Data Mining", AAAI/MIT Press, 1996.

[2] The W3C Technology Stack; "World Wide Web Consortium", Retrieved April 21, 2012.

[3] Gediminas Adomavicius, Alexander Tuzhilin,"Using data mining methods to build customer profiles", IEEE Computer, 34(2), pp. 74-82, Feb 2001.

[4] Ghani, R., A. Fano,"Building Recommender Systems Using a Knowledge Base of Product Semantics", In Proceedings of the Workshop on Recommendation and Personalization in E-Commerce, at the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems, 2002, pp. 11-19.

[5] A. Anitha,"A Web Recommendation Model for e-Commerce Using Web Usage Mining Techniques", Advances in Computational Sciences and Technology, Vol. No. 4, 2010 pp. 507–512.

[6] Buchner et al.,"Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining", SIGMOD Record, 1998, 27(4): pp. 54-61.

[7] Abdelhakim Herrouz, et. al, "Overview of Web Content Mining Tools", The International Journal of Engineering and Science (IJES), Vol. 2, Issue 6, 2013.

[8] Oren Etzioni,"The World Wide Web: Quagmire or gold mine", Communications of the ACM, 39(11), pp. 65-68, 1996.

[9] Jaideep Srivastava et al,"Web usage mining: Discovery and applications of usage patterns from web data", SIGKDD Explorations, 1(2), pp. 12–23, 2000

[10] S. K. Pani,"Web Usage Mining: A Survey on Pattern Extraction from Web Logs", International Journal of Instrumentation, Control and Automation (IJICA), Vol. 1, Issue 1, 2011

[11] Raymond Kosala, Hendrik Blockeel,"Web mining research: A survey", SIGKDD Explorations, pp. 95-104, July 2000.

[12] Arvind Kumar Sharma, et. al,"Exploration of Efficient Methodologies for the Improvement in Web Mining Techniques: A Survey", International Journal of Research in IT & Management, Vol. 1, Issue 3, July 2011.

[13] Agrawal R. et. al,"Mining association rules between sets of items in large databases", In Proceedings ACM SIGMOD International Conference on Management of Data, Vol. 22, No. 2, of SIGMOD Record, Washington, pp. 207–216, 1993.

[14] [Online] Available: http://www.weblogexpert.com

[15] [Online] Available: http://www.123loganalyzer.com

[16] [Online] Available: http://www.dtreg.com/index.html

[17] Parul Agarwal, et. al,"Issues, Challenges and Tools of Clustering Algorithms", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May 2011.

[18] [Online] Available: http://glaros.dtc.umn.edu/gkhome/views/cluto.

[19] Ying Zhao, George Karypis,"Criterion functions for document clustering: Experiments and Analysis", Technical Report TR #01–40, Department of Computer Science, University of Minnesota, Minneapolis, MN, 2001.

[20] Y. Zhao, G. Karypis,"Evaluation of hierarchical clustering algorithms for document datasets", In CIKM, 2002.

[21] [Online] Available: http://www.clustan.com/clustan_package.html

[22] [Online] Available: http://www.gnu.org/software/octave/docs.html

[23] [Online] Available: http://people.sc.fsu.edu/~jburkardt/f_src/spaeth2/spaeth2.html

[24] [Online] Available: http://www.cs.waikato.ac.nz/ml/weka/

[25] Arvind K Sharma, P.C. Gupta, "Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data mining", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Vol. 1, Issue 8, October, 2012.

[26] Screen-scraper, [Online] Available: http://www.screen-scraper.com

[27] [Online] Available: http://www.automationanywhere.com

[28] Zhang, Q., Segall, R.S.,"Web Mining: A Survey of Current Research, Techniques, and Software", International Journal of Information Technology & Decision Making. Vol. 7, No. 4, pp. 683-720. World Scientific Publishing Company (2008).

[29] [Online] Available: http://www.newprosoft.com/web-content-extractor.htm

[30] [Online] Available: http://www.mozenda.com/web-mining-software.

Dr. Arvind K Sharma holds PhD degree in Computer Science. He has more than 14 years of work experience in academic field. He has published more than 29 Papers in various National, International Journals and Conferences. He has authored and co-authored almost 5 books. He has visited Thailand and Dubai for attending International Conferences. He has participated as Speaker and Keynote Speaker in many National and International Conferences. He is a Member of numerous academic and professional bodies i.e. IEEE, WASET, IEDRC, IAENG Hong Kong, IACSIT Singapore, UACEE UK, ACM, New York. He is a Member of Technical Advisory Committee of many International Conferences in India and abroad. He is also Editorial Board Member and Reviewer of several National and International Journals. His area of interest includes Web Usage Mining, Web Intelligence Applications, Web Data Mining, Big Data Analytics and Machine Learning Tools.

Dr. Anubhav Kumar has received Ph.D degree in Computer Science Engineering from the School of Computer and System Sciences, Jaipur National University, Jaipur. He has over 8+ years of teaching experience and authored, co-authored almost 33 research papers in National, International Journals & Conferences. His current area of research includes ERP, KM, Web Usage Mining, 3D Animation. He is a Senior Member of numerous academic and professional bodies such as: IEEE, WASET, IAENG Hong Kong, IACSIT Singapore, UACEE UK, Association for Computing Machinery Inc. (ACM), New York. He is also Reviewer and Editorial Board Member of many International Journals such as IJRECE, IJCAT,IJCDS, IJMR, IJMIT & IJCT. Besides it, he is guiding a few numbers of M.Tech & Ph.D Scholars in the area of Computer Science Engineering.