Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 4, April 2014, pg.825 – 831

REVIEW ARTICLE

A Preliminary Review on Web Usage Mining: A Web Mining Technique

Snehal R. Kawalkar[1], Dr. Pravin P. Karde[2]

[1,2]Department of CS & IT, H.V.P.M. College of Engineering & Technology, Amravati University, India

[1] kawalkar.snehal@gmail.com; [2] p_karde@rediffmail.com

Abstract— Web Mining is that area of Data Mining which deals with the extraction of interesting knowledge from the World Wide Web. There are three important areas in Web Mining: web content mining, web usage mining and web structure mining. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web Usage Mining is that part of Web Mining which deals with the extraction of knowledge from server log files; source data mainly consist of the textual logs that are collected when users access Web servers and might be represented in standard formats, typical applications are those based on user modelling techniques, such as Web personalization, adaptive Web sites, and user modelling. Web Structure Mining mines the structure of hyperlinks within the web itself. This paper focuses on the concept of web usage mining WUM area which is a process of applying data mining techniques to discover interesting patterns from web usage data.

Keywords— Data pre-processing; Web Log; Web mining; Web personalization; Web Usage Mining

I. INTRODUCTION

Web mining can be considered as the applications of the general data mining techniques to the Web. However, the intrinsic properties of the Web make us have to tailor and extend the traditional methodologies considerably. Firstly, even though Web contains huge volume of data, it is distributed on the internet. Before mining, we need to gather the Web document together. Secondly, Web pages are semi-structured, in order for easy processing documents should be extracted and represented into some format. Thirdly, Web information tends to be of diversity in meaning, training or testing data set should be large enough. Even though the difficulties above, the Web also provides other ways to support mining, for example, the links among Web pages are important resource to be used [14].

The goal in web mining is to discover and retrieve useful and interesting patterns from a large dataset. The source data for web mining contains various information sources in different formats. Web usage mining (WUM) is a new research area which can be defined as a process of applying data mining techniques to discover interesting patterns from web usage data [10]. Web usage mining provides information for better understanding of server needs and web domain design requirements of web-based applications [9]. Web usage data contains information about the identity or origin of web users with their browsing behaviours in a web domain. Web pre-fetching, link prediction, site reorganization and web personalization are common applications of WUM. Aim of Web content mining is to extract information relating to the website page contents. It extracts or mines useful information or knowledge from Web page contents [15].

II. WEB MINING

Web mining uses many data mining techniques; it is not purely an application of traditional data mining due to the heterogeneity and semi-structured or unstructured nature of the Web data. Web mining process is similar to the data mining process. The difference is usually in the data collection. In traditional data mining, the data is often already collected and stored in a data warehouse. For Web mining, data collection can be a substantial task, especially for Web structure and content mining, which involves crawling a large number of target Web pages. Once the data is collected, we go through the same three-step process: data pre-processing, Web data mining and post-processing. However, the techniques used for each step can be quite different from those used in traditional data mining [4].

Categories of web mining: Web mining is categorized under three categories as shown in Fig. 1:

1) Web Content Mining: Web content mining is a process of extracting information from texts, images and other contents.

2) Web Usage Mining: Web Usage Mining is a process of extracting information from how to use web sites.

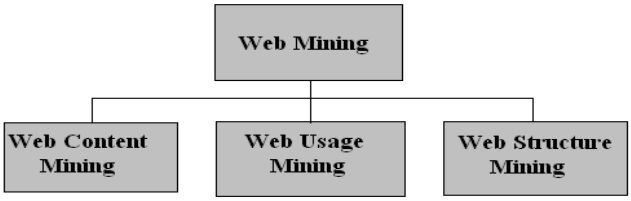

Fig. 1: Categories of Web Mining

3) Web Structure Mining: Web Structure Mining is a process of extracting information from linkages of web pages.

III. WEB USAGE MINING

Web usage mining is an important research area due to following reasons [3]:

a) Web usage mining can be used for personalization for a user .This can be done by keeping track of previously accessed pages of a user .These pages can be used to identify the typical behavior of the user and to make prediction about desired pages.

b) To identify the needed links to improve the overall performance of future accesses, frequent access behavior for the users can be used. Prefetching and caching policies can be made on the basis of frequently accessed pages to improve latency time.

c) To improve the actual design of web pages and for making other Modifications to a Web site, common access behaviors of the users can be used.

d) Usage patterns can be used for business intelligence in order to improve sales and advertisement by providing product recommendations.

A. *Web Log Format*

A web server log file contains requests made to the web server, recorded in chronological order. The most popular log file formats are the Common Log Format (CLF) and the extended CLF. A common log format file is created by the web server to keep track of the requests that occur on a web site [5], [8].

B. *Approach of Web Usage Mining*

The web usage mining typically [5], [8] include the following several steps: data collection, data preprocessing, knowledge discovery and pattern analysis.

1. Data collection: Data Collection is the first step in web usage mining process. It consists of gathering the relevant web data. Data source can be collected at the server-side, client-side, proxy servers, or obtain from an organization's database, which contains business data or consolidated Web data.

Server level collection collects client requests and stored in the server as web logs. Web server logs are plain text that is independent from server platform.

Cookies are unique ID generated by the web server for individual client browsers and it automatically tracks the site visitors. When the user visits next time the request is send back to the web server along with ID. However if the user wishes for privacy and security, they can disable the browser option for accepting cookies.

Explicit User Input data is collected through registration forms and provides important personal and demographic information and preferences. However, this data is not reliable since there are chances of incorrect data or users neglect those sites.

Client Side Collection is advantageous than server side since it overcomes both the caching and session identification problems. Browsers are modified to record the browsing behaviour. Remote agents like Java Applets are used to collect user browsing information. Java applets may generate some additional overhead especially when they are loaded for the first time. But users are to be convinced to use modified browser. Along with log files intentional browsing data from client side like "add to my favourites", "copy" is also added for efficient web usage mining.

Proxy level collection is the data collected from intermediate server between browsers and web servers. Proxy caching is used to reduce the loading time of a Web page experienced by users as well as the network traffic load at the server and client sides. Access log from proxy servers are of same format as web server log and it records the web page request and response for the server. Proxy traces may reveal the actual HTTP requests from multiple clients to multiple Web servers. This may serve as a data source for characterizing the browsing behaviour of a group of anonymous users sharing a common proxy server.

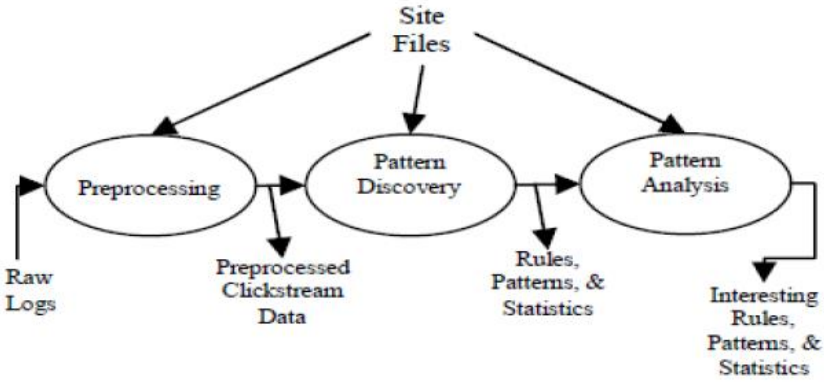

Fig. 2: Web Usage Mining Process

2. Data preprocessing: Some databases are insufficient, inconsistent and including noise. The data pretreatment is to carry on a unification transformation to those databases. The result is that the database will to become integrate and consistent, thus establish the database which may mine. In the data pretreatment work, mainly include data cleaning, user identification, session identification and path completion.

a) Data Cleaning:

The purpose of data cleaning is to eliminate irrelevant items, and these kinds of techniques are of importance for any type of web log analysis not only data mining. According to the purposes of different mining applications, irrelevant records in web access log will be eliminated during data cleaning. Since the target of Web Usage Mining is to get the user's travel patterns, following two kinds of records are unnecessary and should be removed:

- The records of graphics, videos and the format information The records have filename suffixes of GIF, JPEG, CSS, and so on, which can found in the URI field of the every record;
- The records with the failed HTTP status code.

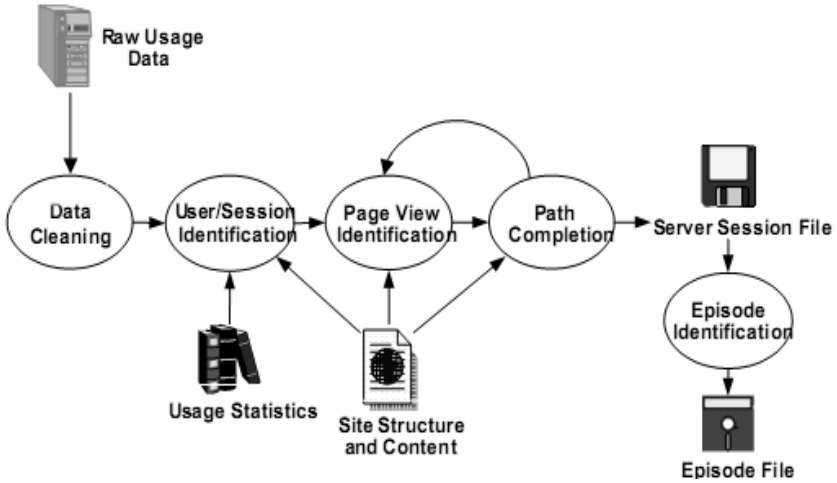

Fig. 3: Preprocessing of Web Usage Data

b) User and Session Identification:

The task of user and session identification is to check out the different user sessions from the original web access log. User' identification is, to identify who access web site and which pages are accessed. The goal of session identification is to divide the page accesses of each user at a time into individual sessions. A session is a series of web pages user browse in a single access. The difficulties to accomplish this step are introduced by using proxy servers, e.g. different users may have same IP address in the log. A referrer-based method is proposed to solve these problems in this study.

c) Path completion

Another critical step in data preprocessing is path completion. There are some reasons that result in path's incompletion, for instance, local cache, agent cache, "post" technique and browser's "back" button can result in some important accesses not recorded in the access log file, and the number of Uniform Resource Locators (URL) recorded in log may be less than the real one. Using the local caching and proxy servers also produces the difficulties for path completion because users can access the pages in the local caching or the proxy servers caching without leaving any record in server's access log. As a result, the user access paths are incompletely preserved in the web access log. To discover user's travel pattern, the missing pages in the user access path should be appended. The purpose of the path completion is to accomplish this task. The better results of data pre-processing, we will improve the mined patterns' quality and save algorithm's running time. It is especially important to web log files, in respect that the structure of web log files are not the same as the data in database or data warehouse. They are not structured and complete due to various causations. So it is especially necessary to preprocess web log files in web usage mining. Through data preprocessing, web log can be transformed into another data structure, which is easy to be mined.

3. Knowledge Discovery: Use statistical method to carry on the analysis and mine the pretreated data. We may discover the user or the user community's interests then construct interest model. At present the usually used machine learning methods mainly have clustering, classifying, the relation discovery and the order model discovery. Each method has its own excellence and shortcomings, but the quite effective method mainly is classifying and clustering at the present.

Classification is the task of mapping a data item into one of several predefined classes. In the web domain, one is interested in the developing a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of a given class or category. Classification can be done by using supervised inductive learning algorithm such as decision tree classifier, naive Bayesian classifier, k- nearest neighbor classifier, Support Vector Machine etc.

Clustering is another data mining technique which is regarded as unsupervised learning process since there are no predefined classes. Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects, so that objects within the same cluster must be similar to some extent, also they should be dissimilar to those objects in other clusters. Use of clustering in web usage mining is to group together similar sessions of users.

For example :a departmental store ,on the basis of different needs of customers can provide different items .In this case the store can provide a better service to its customers[3].

4. Pattern analysis: Challenge of [9], [10] pattern analysis is to filter uninteresting information and to visualize and interpret the interesting pattern to users. Visualization assists an analyst to better apprehend navigation patterns and to predicate trends of data. Knowledge about content and structure also contribute to filtering un-useful knowledge. Many web tools provide some objective criteria, supporting and confidence. Such criteria are helpful to manually filter some believed unimportant knowledge. WebViz tool has done some pioneering work in visualizing of access patterns. It displays access pattern of user as directed graph, with nodes representing page of the access pattern and links representing the hyperlinks between pages. Web pages in the web site can be classified in to

Excellent –the web pages with highest hit counts

Medium - the web pages with average hit counts.

Weak --- the web pages with least hit count.

Web users who have visited web sites can be classified as Class A users (Excellent), Class B users (Medium) and Class C users (Weak users).

IV. PERSONALIZATION

Personalization is all about edifice customer loyalty by building a meaningful one-to-one relationship by understanding the needs of each individual and satisfying a objective that efficiently and knowledgeably addresses each individual's need in a given perspective. Personalization requires implicitly or explicitly collecting visitor information and leveraging that knowledge in your content delivery framework to manipulate

what information you present to your users and how you present it. Personalization for its own sake has the potential to increase the complexity of site interface and drive inefficiency into the architecture. It is the capability to customize customer communication based on knowledge preferences and behaviors at the time of interaction [4]. Web personalization process includes:
Personalization techniques can be divided in three parts:

A. *Content-Based Filtering*

In this, the user model includes information about the content of items of interest- whether these are web pages, movies, music, or anything else. Using these items as a basis, the technique identifies similar items that are returned as recommendations. One of the limitations when using content-based techniques is that it leads to over - specialization.

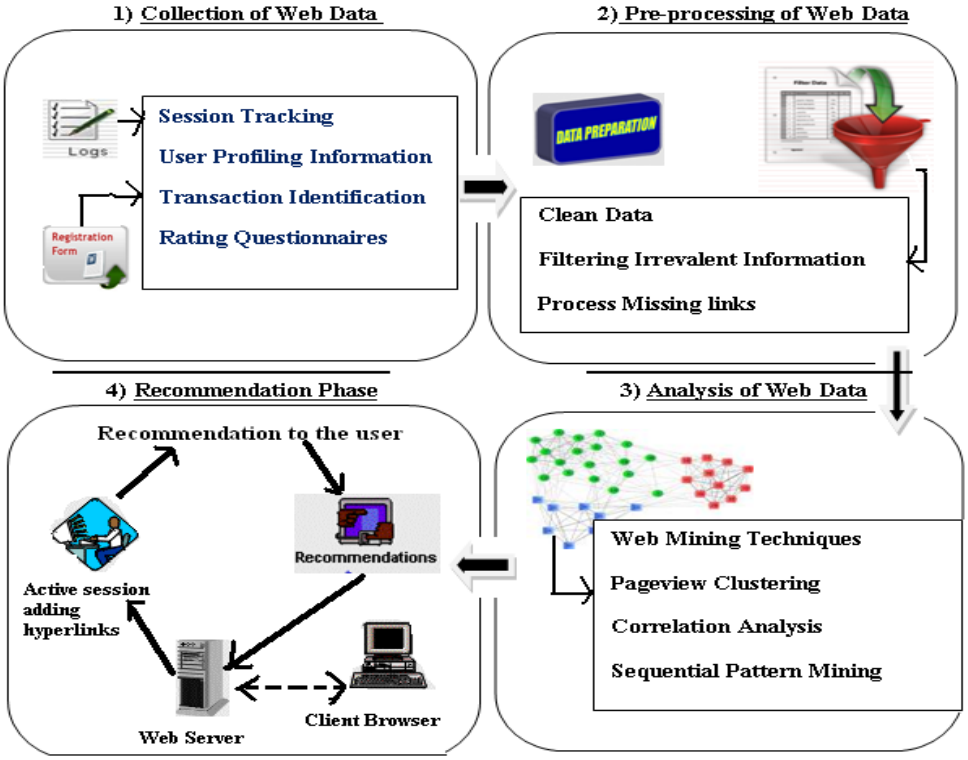

Fig. 4: Web Personalization Process

B. *Collaborative -Based filtering*

In social or collaborative filtering, the system constructs rating profiles of its users, locates other users with similar rating profiles and returns items that the similar users rated highly. Scalability is a problem because computation grows linearly with the number of users and items. The advantage of social filtering, compared to content-based techniques, is that the pool from which recommendations originate is not restricted to items for which the active user has demonstrated interest. The pool will also include items that other users, users that are in some respect similar, have rated highly. This can prove to be instrumental in enhancing the user‘s model: social filtering systems give the user the opportunity to explore new topics and items.

C. *Rule-Based filtering*

It selects only the appropriate service in formations by comparing the query result produced from the Search Manager (Search Module) which is the user interface of the search engine and the rule of the user fetched from the user profile registry

V. RELATED WORK

Tiwari and Renu Tilwani [1] focused on preprocessing approach of Web Usage Data and concept of Web Mining. Jose M. Domenech and Javier Lorenzo [2] presented a tool for Web Usage Mining and implementation was done in Java and making use of the Weka inducers which allow to test new model induction algorithms. Gopal Pandey, Swati Patel, Vidhu Singhal, Akshay Kansara [4] presented a survey on a personalized collaborative filtering method combining the association rule mining focusing on the problems and the solution. Maryam Jafari1, Shahram Jamali, and Farzad Soleymani Sabzchi [6] presented PD-FARM (Pattern Discovery

based on Fuzzy Association Rule Mining) algorithm to extract the web usage patterns, based on Fuzzy Association Rule Mining (FARM). Fuzzy Frequent Pattern-Growth (FFP-Growth) algorithm is used to FARM.

Chhavi Rana [7] focused on Web Usage Mining Research Tools. Dr.D.Suresh Babu, SK.Abdul Nabi, Mohd.Anwar Ali, and Y.Raju focused [8] on Web Usage Mining Process and proposed frame work "Online Miner "seems to work well for developing prediction models to analyze the web traffic volume. Mrs Geeta R.B, Prof. Shashikumar G.Totad, and Dr. Prasad Reddy [10] described importance of web usage mining and its relationship with web structure mining.

V. Shanmuga Priya, S. Sakthivel [11] proposed a new method for web data extraction. It has three phases. In the first phase list of web documents are selected, second phase documents are preprocessed, in the final phase results are presented to users. Abdul-Aziz Rashid Al-Azmi [13] Mining tools such as data mining, text mining, and web mining are used to find hidden knowledge in large databases or the Internet. Mining tools are automated software tools used to achieve business intelligence by finding hidden relations, and predicting future events from vast amounts of data.

VI. APPLICATION OF WEB USAGE MINING

The results produced by the mining of web logs can use for various purposes [3]:

1) To personalize the delivery of web content.

2) To improve user navigation through Prefetching and caching.

3) To improve web design; or in e-commerce sites.

4) To improve the customer satisfaction.

5) Personalization of Web Content: Web Usage Mining techniques can be used to provide personalized web user experience. For instance, it is possible to anticipate, in real time, the user behavior by comparing the current navigation pattern with typical patterns which were extracted from past web log. In this area, recommendation systems are the most common application; their aim is to recommend interesting links to products which could be interesting to users.

6) Prefetching and Caching: The results produced by Web Usage Mining can be exploited to improve the performance of web servers and web-based applications. Typically, Web Usage Mining can be used to develop proper prefetching and caching strategies so as to reduce the server response time as done in.

7) Support to the Design: Usability is one of the major issues in the design and implementation of web sites. The results produced by Web Usage Mining techniques can provide guidelines for improving the design of web applications.

VII. FUTURE

There are a number of issues in preprocessing of log data. Volume of requests in web log in a single log file is the first challenge. Analysing web user access log files helps to understand the user behaviours in web structure to improve the design of web components and web applications. Log includes entries of document traversal, file retrieval and unsuccessful web events among many others that are organized according to the date and time. It is important to eliminate the irrelevant data. So cleaning is done to speed up analysis as it reduces the number of records and increases the quality of the results in the analysis stage. Efforts in this data to find accurate sessions are likely to be the most fruitful in the creation of much effective web usage mining and personalization systems. By following data preparation steps, it is very easier to generate rules which identify directories for website improvement. More research can be done in preprocessing stages to clean raw log files, and to identify users and to construct accurate sessions.

VIII. CONCLUSIONS

Web sites are one of the most important tools for advertisements in international area for universities and other foundation. The quality of a website can be evaluated by analysing user accesses of the website. To know the quality of a web site user accesses are to be evaluated by web usage mining. The results of mining can be used to improve the website design and increase satisfaction which helps in various applications. The goal in web mining is to discover and retrieve useful and interesting patterns from a large dataset. An important research area in Web mining is Web usage mining which focuses on the discovery of interesting patterns in the browsing and navigation data of Web users. This paper presents an overview of web usage mining which mines the log data stored in the web server and framework for web personalization expert based on web mining.

ACKNOWLEDGMENT

The author would like to thank Dr. P. P. Karde and Prof. R. R. Keole H.V.P.M college of Engineering and Technology, Amravati University, India for his support and help during this Paper.

REFERENCES

[1] Sanjay Tiwari, and Renu Tilwani, "Web Mining and Pre-processing of Web Usage Data," International Journal of Advanced Research in Computer Science and Software Engineering, Vol 3, Issue 10, October 2013.

[2] Jose M. Domenech, and Javier Lorenzo, "A Tool for Web Usage Mining" 8th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'07), 16-19 December, 2007, Birmingham, UK.

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

[4] Gopal Pandey, Swati Patel, Vidhu Singhal, and Akshay Kansara, "A Process Oriented Perception of Personalization Techniques in Web Mining" International Journal of Science and Modern Engineering (IJISME) ISSN: 2319-6386, Volume-1, Issue-2, January 2013.

[5] Arun Singh, Avinav Pathak, and Dheeraj Sharma, "Web Usage Mining: Discovery Of Mined Data Patterns and their Applications, International Journal of Computer Science and Management Research Vol 2 Issue 5 May 2013 ISSN 2278-733X.

[6] Maryam Jafari1, Shahram Jamali, and Farzad Soleymani Sabzchi," Discovering Users` Access Patterns for Web Usage Mining from Web Log Files" Journal of Advances in Computer Research Quarterly ISSN: 2008-6148 Sari Branch, Islamic Azad University, Sari, I.R.Iran (Vol. 4, No. 3, August 2013), Pages: 25-32.

[7] Chhavi Rana, "A Study of Web Usage Mining Research Tools" Int. J. Advanced Networking and Applications Volume:03 Issue:06 Pages:1422-1429 (2012) ISSN : 0975-0290.

[8] Dr.D.Suresh Babu, SK.Abdul Nabi, Mohd.Anwar Ali , and Y.Raju, "Web Usage Mining: A Research Concept of Web Mining" International Journal of Computer Science and Information Technologies, Vol. 2 (5) , 2011, 2390-2393, ISSN: 0975-9646.

[9] D.Suresh Babu, P.Sathish, J.Ashok, "Fusion of Web Structure Mining and Web Usage Mining" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (3) , 2011, 965-967.

[10] Mrs Geeta R.B, Prof. Shashikumar G.Totad, and Dr. Prasad Reddy PVGD, "Amalgamation of Web Usage Mining and Web Structure Mining" International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009.

[11] V. Shanmuga Priya,and S. Sakthivel, "AN IMPLEMENTATION OF WEB PERSONALIZATION USING WEB MINING TECHNIQUES" International Journal of Computer Science and Mobile Computing, Vol. 2, Issue. 6, June 2013, pg.145 – 150

[12] V.Chitraa, and Dr. Antony Selvdoss Davamani, "A Survey on Preprocessing Methods for Web Usage Data" International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010.

[13] Abdul-Aziz Rashid Al-Azmi, "DATA, TEXT, AND WEB MINING FOR BUSINESS INTELLIGENCE: A SURVEY" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.2, March 2013.

[14] Miguel Gomes da Costa Júnior Zhiguo Gong, "Web Structure Mining: An Introduction" Proceedings of the 2005 IEEE International Conference on Information Acquisition June 27 - July 3, 2005, Hong Kong and Macau, China

[15] Abdelhakim Herrouz, Chabane Khentout, and Mahieddine Djoudi, "Overview of Web Content Mining Tools" The International Journal of Engineering And Science (IJES) Volume 2, Issue 6, Pages, 2013, ISSN: 2319 – 1813 ISBN: 2319 – 1805

